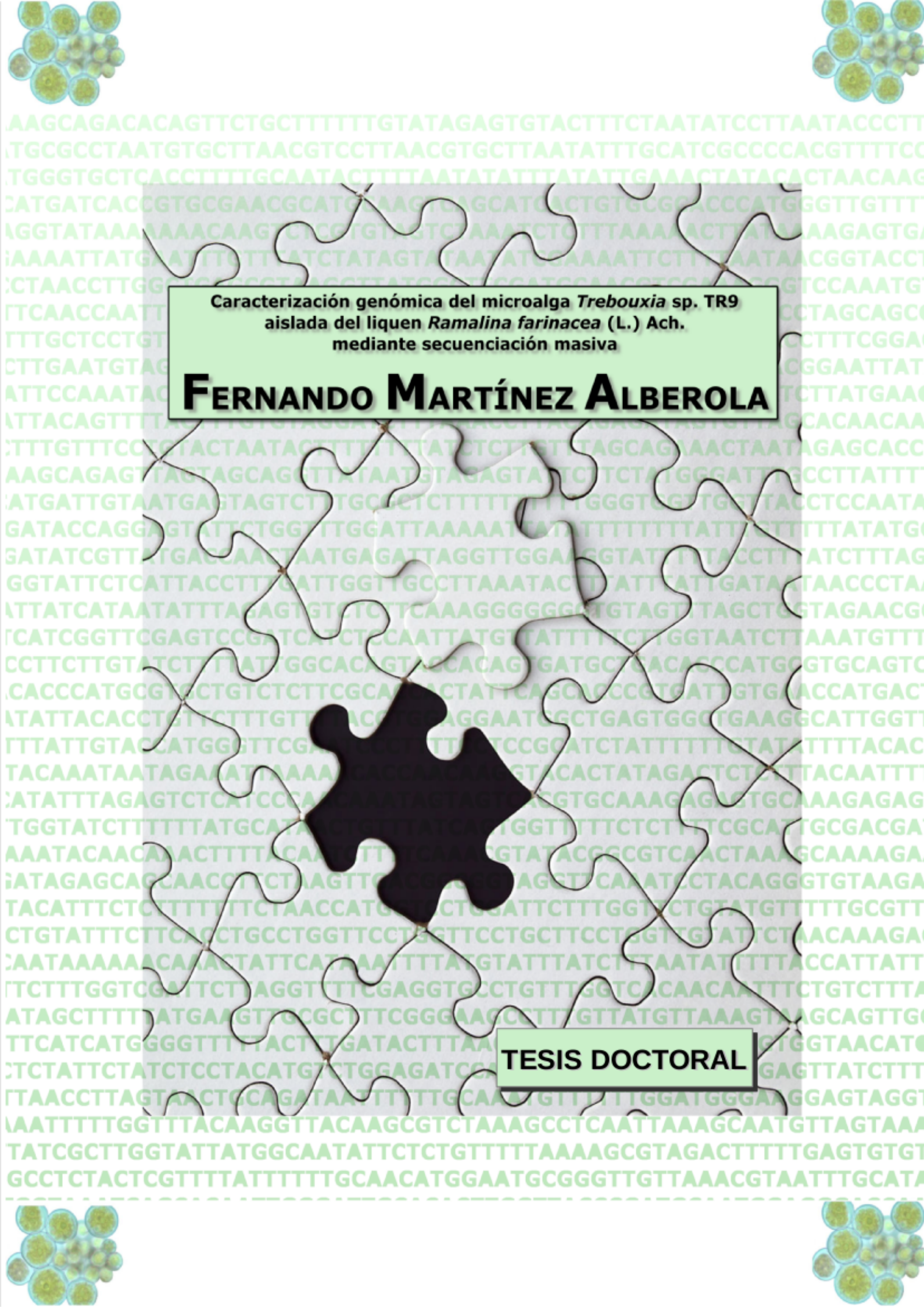


# Caracterización genómica del microalga *Trebouxia* sp. TR9 aislada del liquen *Ramalina farinacea* (L.) Ach. mediante secuenciación masiva

FERNANDO MARTÍNEZ ALBEROLA

TESIS DOCTORAL

VNIVERSITAT DE VALÈNCIA

Instituto Universitario Cavanilles de Biodiversidad y Biología Evolutiva

Facultad de Ciencias Biológicas

# Caracterización genómica del microalga *Trebouxia* sp. TR9 aislada del liquen *Ramalina farinacea* (L.) Ach. mediante secuenciación masiva

FERNANDO MARTÍNEZ ALBEROLA

Programa de Doctorado en Biodiversidad y Biología Evolutiva

Directora: Eva Barreno Rodríguez

Codirectora: Eva María del Campo López

Valencia 2015

La Dra. Eva Barreno Rodríguez, Catedrática de Botánica de la Universitat de València, y la Dra. Eva María del Campo López, Profesora Contratada Doctora del Departamento de Ciencias de la Vida de la Universidad de Alcalá

Certifican:

Que la memoria presentada por Fernando Martínez Alberola con título "Caracterización genómica del microalga *Trebouxia* sp. TR9 aislada del liquen *Ramalina farinacea* (L.) Ach. mediante secuenciación masiva" corresponde a su tesis doctoral y ha sido realizada bajo su dirección en el Instituto Universitario Cavanilles de Biodiversidad y Biología Evolutiva, autorizando mediante este escrito la presentación de la misma para optar al grado de Doctor en Ciencias Biológicas por la Universitat de València.

Y para que así conste a los efectos oportunos, en cumplimiento de la legislación vigente, firmamos el presente informe en Burjassot, Septiembre de 2015

Fdo.: Eva Barreno Rodríguez, Catedrática de Botánica de la Universitat de València

Fdo.: Dra. Eva María del Campo López, Profesora Contratada Doctora del Departamento de Ciencias de la Vida de la Universidad de Alcalá

Fdo.: Fernando Martínez Alberola

*"Además de la ley de la lucha mutua, existe en la naturaleza también la ley de ayuda mutua, que, para el éxito de la lucha por la vida y, particularmente, para la evolución progresista de las especies, desempeña un papel mucho más importante que la ley de la lucha mutua"*

Piotr Alekséyevich Kropotkin

El Apoyo Mutuo (1902)

# Resumen

Los líquenes son complejas entidades simbióticas originadas por asociaciones cíclicas e interacciones positivas de organismos muy diferentes, un hongo heterotrófico y uno o varios "fotobiontes" autótrofos fotosintéticos, tanto algas verdes unicelulares, como cianobaterias o ambos. Además, comunidades específicas de bacterias aparecen como simbiontes obligados de líquenes. En todos los talos del liquen *Ramalina farinacea* (L.) Ach., coexisten las dos mismas especies de algas *Trebouxia* (*T. jamesii* -TR1- y *Trebouxia* sp. TR9). Esta forma de asociación simbiótica específica y selectiva es interesante como modelo de la coexistencia de dos especies diferentes de *Trebouxia* dentro de un mismo talo liquénico. Esta coexistencia es probable que sea promovida por diferentes contextos fisiológicos. *Trebouxia* sp. TR9 muestra respuestas originales que se inducen frente al estrés abiótico y tiene un mejor rendimiento fisiológico que *T. jamesii*. Estas características pueden ser el reflejo de su mayor capacidad para llevar a cabo ajustes metabólicos clave. Para investigar la base genética de la plasticidad de esta especie de microalga, se ha generado un estudio de las secuencias genómicas de *Trebouxia* sp. TR9 utilizando NGS. El ADN total obtenido de cultivos de *Trebouxia* sp. TR9 aislados del liquen *Ramalina farinacea* fue secuenciado con las tecnologías 454 y Solexa con objeto de analizar su estructura genómica e investigar nuevas secuencias de genes mediante la comparación con las bases de datos existentes de genomas de algas verdes.

Como resultado, se han obtenido las secuencias de los genomas cloroplástico, mitocondrial y nuclear de *Trebouxia* sp. TR9. Los genomas mitocondriales y cloroplásticos han sido anotados de forma manual y se han identificado en ellos un total de 61 y 108 genes, respectivamente. El genoma mitocondrial es de naturaleza circular y tiene un tamaño de 70.070 nt. en algunos de los genes anotados se han localizado 9 intrones de tipo I, algunos de los cuales contienen ORFs que codifican "Homing endonucleases" de la familia LAGLIDADG.

El genoma cloroplástico de *Trebouxia* sp. TR9 es también de naturaleza circular y posee la estructura cuatripartita típica de los cloroplastos de plantas vasculares, las regiones repetidas invertidas o IR incluyen un único gen, el *rbcL*. El tamaño final del genoma cloroplástico obtenido ha sido de 303.323 nt siendo uno de los mayores tamaños conocidos en el contexto de las algas verdes de la división Chlorophyta. En algunos genes, se han identificado 12 intrones de tipo I, 6 de ellos con ORFs que podrían codificar "Homing endonucleasas" de la familia LAGLIDADG.

El genoma nuclear de *Trebouxia* sp. TR9 ha sido ensamblado con el programa Velvet utilizando diferentes tamaños de K-mer (21-133). El ensamblaje más óptimo del genoma nuclear abarcaba 2.626 "contigs" que tenían una longitud total de 59.121.427 nt junto a un tamaño de "contig" N50 y N95 de 142.866 nt y 21.727 nt, respectivamente. Para

comprobar la continuidad del ensamblaje nuclear se ha utilizado el espacio génico calculado con la herramienta CEGMA. El genoma nuclear comprendía un 91 y 97 % del conjunto de 248 CEGs de forma completa y parcial, respectivamente. La anotación del genoma nuclear se basó en la predicción de genes "ab initio" con el programa AUGUSTUS entrenado con los modelos obtenidos anteriormente con el programa CEGMA. Como resultado, se obtuvieron 9.499 posibles modelos génicos, 6.364 fueron anotados con al menos un término de Gene Ontology (GO) y de ellos, 2.249 mostraron un número enzimático asociado. Se han encontrado elementos móviles en el genoma nuclear, la mayoría de ellos son retrotransposones con repeticiones terminales largas (LTR), de los cuales, los más abundantes son de tipo Gypsy/DIRS1 y Ty1/Copia. Se han detectado un total de 10.922 dominios proteicos PFAM presentes en 6.544 modelos proteicos de los que 2.147 y 2.979 son compartidos por todas o por, al menos, una de las especies de microalgas verdes con las que se ha comparado este genoma (*Asterochloris* sp., *Chlorella variabilis*, *Coccomyxa subellipsoidea* y *Chlamydomonas reindhartii*), respectivamente. Además, se han encontrado 19 motivos PFAM específicos de la división Chlorophyta, 6 motivos propios de las algas liquénicas *Trebouxia* sp. TR9 y *Asterochloris* sp y 23 motivos propios de *Trebouxia* sp. TR9. La identificación de proteínas relacionadas con la asimilación de carbono sugiere que *Trebouxia* sp. TR9 puede tener mecanismos de concentración de carbono de tipo C3 y C4/CAM. El estudio de enzimas relacionadas con el metabolismo de carbohidratos presentes en los modelos proteicos de *Trebouxia* sp. TR9 y otras algas de la clase Trebouxiophyceae indica que, entre el 11 y el 12 % de las proteínas totales de estas algas, pertenecían al menos a un clan de las familias de la base de datos "Carbohydrate-Active enZYmes Database" (CAZy). Además, las anotaciones de genes revelaron la existencia de proteínas relacionadas con virus. Los datos obtenidos en esta tesis doctoral no sugieren una reducción dramática del genoma de *Trebouxia* sp. TR9 que pueda relacionarse a su forma de vida simbiótica.

Este trabajo ofrece una visión general de los tres genomas del microalga *Trebouxia* sp. TR9 y de sus principales características y estructuras, que aportan datos inéditos y relevantes para el análisis de las tendencias evolutivas de las Trebouxiophyceae. Así, p. ej., respecto a otras filogenias realizadas en base a genes cloroplásticos, la realizada aquí con las secuencias de las proteínas codificadas por siete genes del genoma mitocondrial de *Trebouxia* sp. TR9 y los de 25 especies de algas verdes, han puesto de manifiesto dos nuevas diferencias fundamentales: (i) la monofilia de la clase Trebouxiophyceae y (ii) la posición de la clase Pedinophyceae más relacionada con las Chlorophyceae y Ulvophyceae.

# Abstract

Lichens are complex symbiotic entities originated by cyclical associations and positive interactions of very different organisms, a heterotrophic fungus and one or several photosynthetic autotrophs "photobionts", either unicellular green algae, cyanobacteria or both. Also, specific bacterial communities remain obligate lichen symbionts. In the lichen thalli of *Ramalina farinacea* (L.) Ach., the same two algal *Trebouxia* species coexist in every thallus (*T. jamesii* -TR1- and *Trebouxia* sp. TR9). This specific and selective form of symbiotic association is interesting as a model of the coexistence of two different *Trebouxia* species within the same lichen thallus. Such coexistence is likely to be promoted by different physiological backgrounds. *Trebouxia* sp. TR9 shows novel inducible responses against abiotic stress and seems to have a better physiological performance than *T. jamesii*. These features may reflect its greater capacity to undertake key metabolic adjustments. To investigate the genetic basis of the physiological plasticity of this microalgal species, a survey of the genomic sequences of *Trebouxia* sp. TR9 by NGS has been generated. The total DNA obtained from cultures of *Trebouxia* sp. TR9, isolated from the lichen *Ramalina farinacea*, was sequenced by 454 and Solexa technologies in order to analyze its genome structure as well as to explore new gene sequences by comparing it with available green algae genome databases.

As a result, the sequences of the chloroplast, mitochondrial and nuclear genomes of *Trebouxia* sp. TR9 were produced. Mitochondrial and chloroplast genomes were annotated manually and a total of 61 and 108 genes, respectively, have been identified. The mitochondrial genome is circular and its size is 70,070 bp, 9 type I introns have been detected in several genes; somo of them contains ORFs codifying “Homing endonucleases” of the LAGLIDADG family.

The chloroplast genome of *Trebouxia* sp. TR9 is also circular and shows the typical quadripartite structure of land plant chloroplasts and the IR, or inverted repeat regions, include a single gene, *rbcL*. The final size of the chloroplast genome was 303,323 pb, this being one of the largest known in the Chlorophyta green algae context. 12 type I introns have been detected; six of them contain ORFs codifying Homing endonucleases of the LAGLIDADG family.

The nuclear genome of *Trebouxia* sp. TR9 has been assembled with Velvet software using different K-mer sizes (21-133). The most optimal assembly comprise 2,626 "contigs" with a total length of 59,121,427 bp, N50 and N95 "contig" length of 142,866 bp and 21,727 bp, respectively. To check the continuity of the nuclear assembly the CEGMA gene space was calculated. The nuclear genome comprised 91 % and 97 % of the total of 248 complete and incomplete CEGs, respectively. The "ab initio" annotation gene prediction was based on AUGUSTUS software trained with the models previously obtained with CEGMA. 9,499 possible gene models were obtained, 6,364 were annotated sho-

wing at least one term of Gene Ontology (GO), and of those 2,249 were associated with an enzyme number. We also found mobile elements in the nuclear genome; most of them were retrotransposons with long terminal repeats (LTR), of which the most abundant were Gypsy / DIRS1 and Ty1 / Copia types. A total of 10,922 PFAM protein domains were present in 6,544 protein models and of those 2,147 and 2,979 were shared by all or at least one of the additional microalgae genomes analyzed (*Asterochloris* sp., *Chlorella variabilis*, *Coccomyxa subellipsoidea* and *Chlamydomonas reindhartii*), respectively. In addition, 19 Chlorophyta specific PFAM motifs, 6 specific motifs of *Trebouxia* sp. TR9 and *Asterochloris* sp lichen algae, and 23 specific motifs of *Trebouxia* sp. TR9 were found. The identification of proteins involved in carbon uptake suggests that *Trebouxia* sp. TR9 may possess carbon concentration mechanisms similar to C3 and C4/CAM. The study of carbohydrate metabolism in the protein enzymes models of *Trebouxia* sp. TR9, and other Trebouxiophyceae algae, point out that between 11 % and 12 % of the total proteins of these algae fitted to at least one clan of the families of the "Carbohydrate-Active Enzymes" (CAZy) database. Furthermore, gene annotations showed virus-related proteins. No dramatic reductions in the genome of *Trebouxia* sp. TR9 associated with its symbiotic way of life are suggested by our data.

This work provides a first glimpse into the three genomes of the microalga *Trebouxia* sp. TR9 and a general overview of their main features and structures which may shed light on the evolutionary trends of the Trebouxiophyceae. For instance, the phylogeny obtained using the sequences codified by seven genes of the mitochondrial genome of *Trebouxia* sp. TR9 and additional ones from 25 species of green algae, revealed two new differences in relation with those obtained using only chloroplast genes: (i) The Trebouxiophyceae are monophyletic and (ii) the Pedinophyceae are more related to Chlorophyceae and Ulvophyceae.

# AGRADECIMIENTOS

Este proyecto ha podido ser llevado a cabo gracias a la colaboración interdisciplinar de muchas personas, a las que debo reconocer su valiosa colaboración:

Quiero expresar mi más sincero agradecimiento a la directora y a la codirectora de esta tesis doctoral, Eva Barreno Rodríguez y Eva María del Campo López, por la confianza que han depositado en mí, así como por sus orientaciones durante el desarrollo de este trabajo, como los medios humanos y materiales que ha puesto a mi disposición. También debo agradecerle el modo en que ha contribuido a mi formación como liquenólogo, como analista bioinformático, como persona y como investigador.

Hago extensivo este agradecimiento a Francisco Gasulla, Carolina Royo, Guillermo Salvá, Sergio Mezquita, Santi Català, Salva Chiva, Ernesto Hinojosa y por supuesto a Arantxa Molíns y Patricia Moya por su inestimable ayuda en las horas de laboratorio, así como por su interés y compañerismo junto a las ideas y sugerencias, de inestimable valor, necesarias para que este trabajo llegase a buen puerto. No pueden faltar en este apartado Francisco García Breijo (UPV) y José Reig Armiñana (Jardí Botànic) gracias a los cuales he aprendido las técnicas anatómicas; a Alejandro Soriano y Josep Sala, de la Unidad de Cálculos del Servei d'Informatica de la Universitat de València, que me han ayudado en el manejo del servidor Lluis Vives; a Vicente Sentandreu y Amaro Martínez, de la Secció de Genómica de la Universitat de València, por su colaboración en las secuenciaciones realizadas y a Leonardo Casano, de la Universidad de Alcalá de Henares, por su ayuda y conocimientos científicos que me ha transmitido sobre el estrés oxidativo y la proteómica.

Quedo en deuda con muchas personas, imposible citarlas a todas sin riesgo de olvidar a algunas, por una ayuda que va desde la solidaridad personal o la pura amistad hasta el acceso a recursos técnicos. En este sentido quiero agradecer su apoyo y colaboración desinteresada y en especial a Isabel Martínez Nieto junto con todos y todas mis amigos y amigas, por el apoyo y amistad prestados hacia mi persona.

Por último, y no menos importante, a mi familia natal, que me han dado la vida, el apoyo, el cariño y la educación necesarios para poder llegar a ser una persona. No importa a donde vaya, ni el tiempo ni la distancia, sé que cuando regrese a casa, me esperarán con los brazos abiertos, así como con los corazones repletos de amor. Y como no, a Inma, mi nueva familia formada con esta maravillosa persona, gracias por todo el apoyo, comprensión y cariño que ha hecho que sea la persona más feliz del mundo. Gracias por ser como son desde el fondo de mi corazón.

A los proyectos de investigación PROMETEO 2008/174, PROMETEO FASEII2013/021, CGL2009-13429-C02-01 y CGL2012-40058-C02-

01 con los que se ha podido financiar el capital humano y técnico para la realización de la presente tesis doctoral.

Como buen manchego que soy, "muchísmas" gracias.

# ÍNDICE GENERAL

# 1

# INTRODUCCIÓN GENERAL

La fotosíntesis oxigénica, el proceso químico por el que la conversión de dióxido de carbono en compuestos orgánicos y oxígeno es alimentado por la energía luminosa, evolucionó de un ancestro anoxigénico de cianobacterias. Existe un consenso general sobre la adquisición de las reacciones fotosintéticas por eucariotas a través de procesos de endosimbiosis. Éstos se llevaron a cabo al unirse dos organismos diferentes, un organismo eucariota heterotrófico que poseía otro orgánulo bioenergético (la mitocondria), junto a los procesos de mitosis, meiosis y otras características propias de eucariotas, que capturó una cianobacteria fotoautotrófica que se integró y finalmente se convirtió en un plasto (Archibald, 2011). La diversificación a partir de este primer eucariota portador de cloroplastos primarios dio lugar tanto al linaje verde, como a las algas rojas y a los glaucófitos. A partir de este punto de partida, la fotosíntesis se extendió en diversos protistas eucarióticos mediante endosimbiosis secundarias y terciarias, dando lugar a los linajes de chloraracniofitos, euglenidos y dinoflagelados "verdes" (Endosimbiosis secundarias) y a los linajes de criptofitos, haptofitos, stramenopiles fotosintéticos, y dinoflagelados (Endosimbiosis terciarias) Leliaert *et al.* (2012). Este grupo de eucariotas, dentro del linaje de Archaeplastida, comprende a organismos que son muy diversos tanto desde el punto de vista filogenético, como morfológico, bioquímico o ecológico. La gran mayoría de estas complejidades son remarcadas por el hecho de ser historias evolutivas complejas debidas a diferentes procesos de endosimbiosis (Figura 1), aunque sus cloroplastos tienen historias evolutivas comunes, los organismos hospedadores que los contienen no (Keeling, 2013).

Los organismos que comprenden el clado de plantas verdes (Viridiplantae) incluyen a las plantas terrestres embriofitas y las algas verdes clorofitas. La monofila de este grupo está bien establecida y análisis filogenéticos junto al escaso registro fósil han calibrado su origen en torno los 700 - 1500 millones de años (Leliaert *et al.* , 2012). Las algas verdes, que dieron origen a las plantas terrestres han tenido una gran importancia ecológica en la evolución de la vida sobre el planeta Tierra, este evento inició el desarrollo de los ecosistemas terrestres y han llevado a cambios ambientales geoquímicos a escala global (De Clerck *et al.* , 2012). Comparten características únicas, sus cloroplastos se ven rodeados por dos membranas, sus tilacoides se agrupan en lamelas y contienen pigmentos clorofílicos a y b junto a otros pigmentos accesorios como los carotenos y las xantófilas. Cuando los pirenoides están presentes, se embeben en la matriz cloroplástica y se ven rodeados de almidón, la principal fuente de reservas polisacáridas. Las paredes celulares, cuando se ven presentes, se componen generalmente de celulosa. Muchas células de algas verdes son flageladas o poseen flagelos en alguna fase de su ciclo vital. Generalmente poseen dos fla-

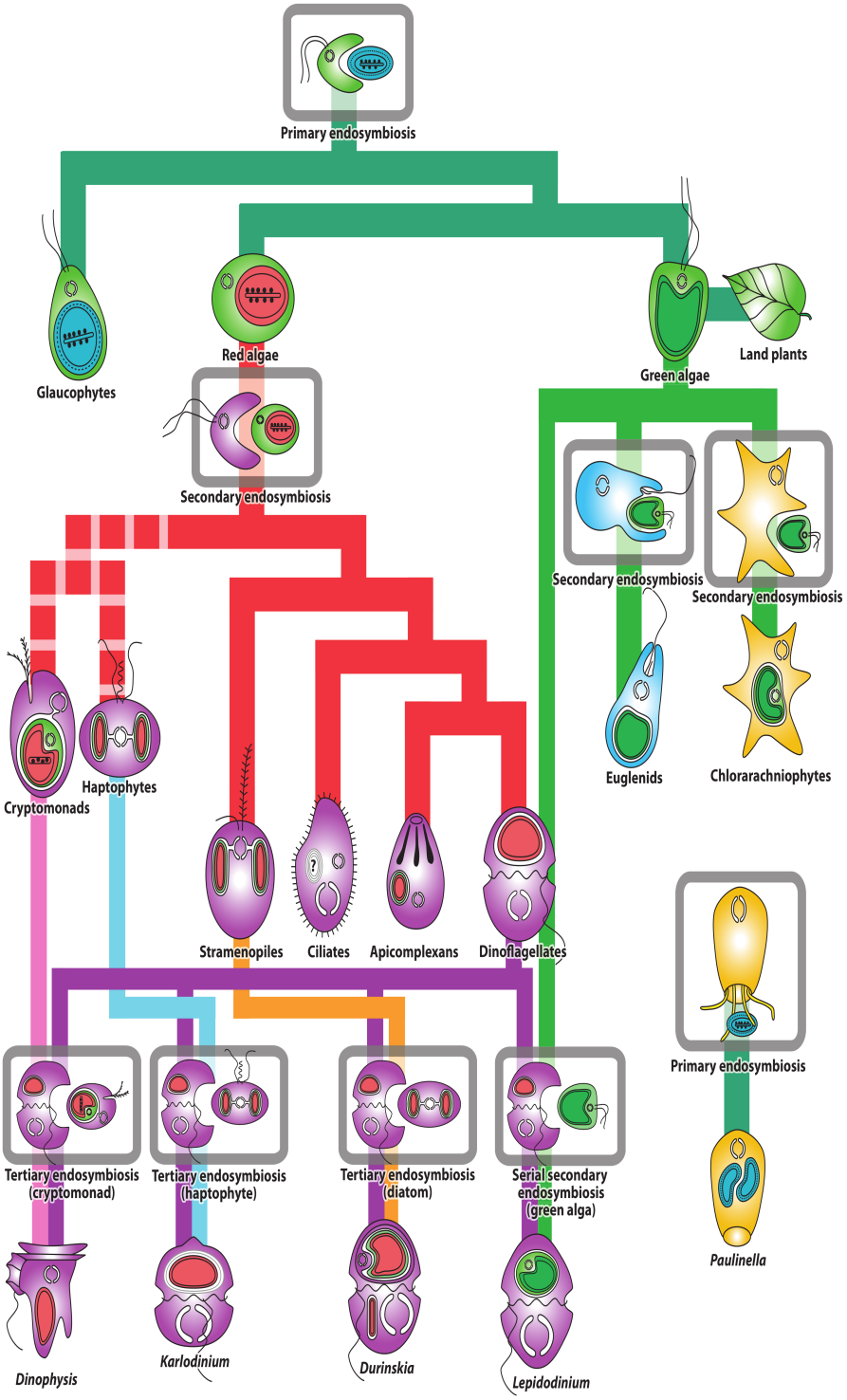



Figura 1: **Esquema de los eventos endosimbióticos en la evolución plastidial.** Cada evento endosimbiótico se encuentra rodeado por un cuadro. Las diferentes líneas unen los linajes procedentes de cada evento de endosimbiosis, las líneas verde y roja muestran los linajes de algas verdes y rojas respectivamente. Imagen tomada de Keeling (2013).

gelos isocontos por célula, de estructura similar pero de dimensiones diferentes. Estudios ultraestructurales del flagelo y el cuerpo basal, los procesos de mitosis y división fueron las primeras características para la clasificación filogenética de las algas verdes (De Clerck *et al.* , 2012; Leliaert *et al.* , 2012). Los datos ultraestructurales sirvieron para una clasificación natural de las algas verdes, pero las relaciones dentro de las diferentes líneas evolutivas no fueron resueltas hasta la aparición de las técnicas filogenéticas moleculares. Puesto que todos los organismos poseen ribosomas, las primeras filogenias moleculares se basaron en la sub-unidad pequeña del ARN ribosomal nuclear (18S), corroborando las clasificaciones basadas en ultraestructura aunque al ser análisis de genes individuales, se obtenía poca resolución. Es por ello - junto a la naturaleza endosimbiótica de los genomas organulares - que en los últimos años se ha optado por construir filogenias de varios genes codificantes concatenados contenidos en los genomas organulares compartidos por las diferentes algas secuenciadas (Wolff *et al.* , 1994; Wakasugi *et al.* , 1997; Turmel *et al.* , 1999b, 2009; De Cambiaire *et al.* , 2007; Brouard *et al.* , 2010; Smith *et al.* , 2011).

Existen dos linajes principales dentro de la división Viridiplantae, clorofitos y estreptofitos. Los clorofitos comprenden la mayoría de especies de algas verdes mientras que los estreptofitos se componen del grupo de algas parafiléticas conocidas como carofíceas y las plantas terrestres embriofitas (Figura 2). Las carofíceas tienen una gama morfológica que oscila entre organismos unicelulares a formas pluricelulares y según numerosos estudios son los seres vivos más próximos a las plantas terrestres, aunque debido al tiempo de divergencia entre estos taxones, junto a la falta de muestreo de suficientes taxones y genes, los estudios filogenéticos realizados posicionan a cada una de las diferentes familias como hermanas de las plantas terrestres Turmel *et al.* (2006); Lemieux *et al.* (2007); Turmel *et al.* (2009); Ruhfel *et al.* (2014).

En la otra rama de la división Viridiplantae se encuentran las algas verdes, éstas a su vez se dividen en dos grupos polifiléticos, las prasinofíceas y los clorofíceas. Las primeras comprenden a un grupo parafilético en la base del árbol filogenético de las clorofitas (Figura 2). La mayoría son unicelulares con diversas morfologías de vida libre (con o sin flagelo), en algunos casos con escamas corporales y que ocupan predominantemente ambientes marinos y en algunos casos de agua dulce. Este grupo de algas contiene a las mamielofíceas, que incluyen a los organismos eucariotas más pequeños conocidos (*Ostreococcus* y *Micromonas*) cuyos genomas han sido secuenciados y son compactos y de pequeño tamaño (Derelle *et al.* , 2006; Worden *et al.* , 2009).

Las algas clorofitas forman un clado diverso tanto a nivel morfológico como ecológico, colonizando ambientes marinos, de aguas dulces y

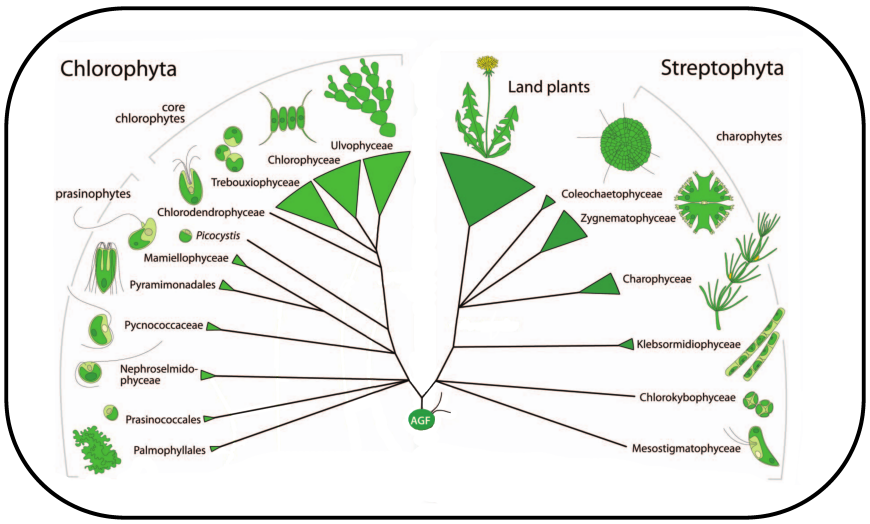



Figura 2: **Filogenia resumen de Viridiplantae.** La topología representa una visión conservativa de las politomías conflictivas que aparecen en diferentes estudios. Todos los linajes parten del primer fenómeno de endosimbiosis llevado a cabo por un flagelado ancestral verde (AGF, Ancestral Green Flagelate). Figura adaptada de Leliaert *et al.* (2012)

algas terrestres. En su base se encuentran las algas clorodendroficeas, marinas y de aguas continentales (Figura 2). Los clados más numerosos y diversos dentro de las clorofitas son los clados Ulvophyceae, Trebouxiophyceae y Chlorophyceae (UTC). Adaptaciones morfológicas y ecológicas del clado monofilético UTC probablemente les han permitido la radiación de las Trebouxiophyceae y Chlorophyceae en hábitats de aguas dulces y aéreos mientras que las Ulvophyceae se han especializado en ambientes litorales marinos. Además, muchas especies del clado UTC se han especializado como organismos simbiontes de líquenes, protistas, invertebrados y plantas; otras han evolucionado de vida libre o parásitas (De Clerck *et al.* , 2012; Leliaert *et al.* , 2012).

Un muestreo más amplio de taxones y/o análisis de grandes conjuntos de datos, tales como genomas / transcriptomas completos, probablemente serán necesarios para arrojar más luz sobre las relaciones filogenéticas del grupo de las plantas (Ruhfel *et al.* , 2014).

## 1.2 FOTOBIONTES LIQUÉNICOS

### 1.2.1 *Simbiosis liquénica*

Los líquenes son organismos complejos resultado de asociaciones simbióticas entre un hongo (micobionte), uno o varios fotobiontes (algas y/o cianobacterias) y consorcios bacterianos no fotosintéticos.

Presentan una organización biológica especial a través del proceso de simbiogénesis, dando como resultado una entidad única u holobionte (Margulis & Barreno, 2003). Recientemente se ha propuesto que los líquenes son sistemas más complejos de lo que se creía, al incluir consorcios bacterianos no fotosintéticos. Las simbiosis liquénicas son procesos cíclicos comparables a los corales, donde la unión sinérgica de las partes de esta asociación no es una simple mezcla, ya que como resultado se forma una estructura- característica y única de los líquenes -en la que el hongo origina un talo que envuelve al fotobionte. El talo está formado por un córtex y una capa medular (constituidos por tejido fúngico), y por una capa algal en la que las células del fotobionte quedan englobadas por las hifas fúngicas.

Desde el momento en que se propuso por primera vez que los líquenes eran organismos duales (Schwendener, 1869), se ha planteado una fuerte controversia, que se ha mantenido a lo largo de la historia de la liquenología, acerca de si la asociación liquénica se puede considerar mutualista o si es un parasitismo. Autores como Schwendener (1869), o como Ahmadjian & Jacobs (1981) han considerado que las células del fotobionte son "víctimas" del micobionte, sosteniendo que la simbiosis liquénica es un parasitismo controlado, y no una relación mutualista. Sin embargo muchos biólogos consideran los líquenes como uno de los mejores ejemplos de mutualismo dada la ventaja adaptativa que les proporciona el micobionte a los productores primarios de la asociación (fotobiontes), el poder colonizar nichos únicos en los que no podrían desarrollarse. Más recientemente Kappen (1994) recalcó la conveniencia de considerar definiciones que traten al liquen como un sistema de autótrofos y heterótrofos. El fotobionte como productor primario de la asociación proporciona carbohidratos al micobionte y éste le surte de agua, elementos minerales y protección frente a la radiación luminosa.

La definición aprobada en 1982 por la Asociación Internacional de Liquenología decía que "un liquen es una asociación entre un hongo y un simbionte fotosintético de la cual resulta un talo estable de estructura específica". Posteriormente, Hawksworth (1988), definió el liquen como "una asociación estable y autónoma entre un micobionte y un fotobionte en la que el micobionte actúa como exhabitante". Sin embargo, Ahmadjian (1993), indicó que en esta definición existían dos términos confusos. El primero es que esta definición no reconoce la remarcable transformación que experimenta el micobionte cuando entra en contacto con el fotobionte (formación del talo). Y en segundo lugar, en ciertos líquenes las células fúngicas penetran en el interior de las células algales a través de los haustorios. En este caso el micobionte se podría considerar inhabitante y no exhabitante, además, también indicó que un liquen es "una asociación entre un hongo, frecuentemente un ascomicete aunque en algunos casos también un basidiomicete o un deuteromicete, y uno o más componentes fotosin-

téticos, generalmente un alga verde o una cianobacteria. En todos los líquenes el hongo forma un talo o un estroma liquenizado que puede contener compuestos secundarios únicos en estos organismos". En este sentido, los talos de los líquenes también pueden interpretarse como microecosistemas, tanto por su propio funcionamiento interno y las relaciones de materia y energía de sus componentes biológicos, como en el funcionamiento global, ya que aceleran los procesos de meteorización de sus sustratos y ponen en circulación muchos iones, sustancias diversas y, en algunos ecosistemas, contribuyen sustancialmente a la fijación del nitrógeno atmosférico.

Los talos liquénicos presentan, frente a los simbiontes aislados, gran originalidad morfológica, fisiológica, adaptativa, de modos de vida y también en cuanto a su modo de reproducción, es decir, constituyen innovaciones simbioespecíficas. Como bien señalan Margulis & Barreno (2003) los líquenes son un buen ejemplo de cómo la integración cíclica de los simbiontes que participan proporciona el potencial de nuevas y distintas relaciones entre organismos y puede ser un mecanismo de innovación evolutiva con efectos morfogenéticos (simbiogénesis); es decir, estar en el origen de nuevas entidades con propiedades emergentes las cuales no son el resultado de la suma linear de las partes. Aunque no hay una definición satisfactoria de liquen, y existe una importante controversia acerca del tipo de relación que se establece entre los simbiontes, lo que sí está claro es que la estabilidad de esta asociación, junto con la naturaleza poiquilohídrica de los líquenes determinan su gran amplitud ecológica.

A lo largo de la evolución las células vegetales han ido desarrollando mecanismos que les permiten evitar o revertir los daños provocados por especies reactivas de oxígeno (ROS). Dichos mecanismos han sido extensamente estudiados en plantas vasculares, e incluyen actividades enzimáticas como glutatión reductasa, superóxido dismutasas, peroxidasas (incluida ascorbato peroxidasa), peroxirredoxinas, catalasa, y antioxidantes no enzimáticos como ascorbato, glutatión, tocoferol y carotenoides (Noctor & Foyer, 1998; Alscher *et al.* , 2002; Asada, 2006; Halliwell, 2006). Otro mecanismo importante bajo condiciones de estrés es la disipación no-fotoquímica de la energía (NPQ), que transforma en calor el exceso de energía luminosa que no puede ser utilizada en fotosíntesis y que podría conducir a la formación de ROS. En las plantas esa disipación del exceso de energía luminosa está asociada con la acidificación del lumen de los tilacoides, el ciclo de las xantófilas y la acción de la proteína PsbS asociada al fotosistema II (Niyogi *et al.* , 2005). Aparentemente, en organismos autótrofos homeohidros el balance entre conservación y disipación de la energía luminosa está desplazado hacia los procesos de conservación de la energía mientras que, en organismos poiquilohidros, son dominantes los procesos de disipación (Heber *et al.* , 2006). Los líquenes son organismos poiquilohídricos capaces de vivir (metabólicamente inactivos)

durante períodos de tiempo prolongados con un contenido hídrico en sus talos igual o inferior al 10 % de su peso seco (Pintado *et al.* , 2005; Sancho *et al.* , 2007). En el experimento realizado por De La Torre *et al.* (2010), donde diferentes muestras liquénicas fueron lanzadas a la órbita terrestre a unos 300 km de altitud en una cápsula espacial y volvieron a la Tierra tras 10 días (Se completaron 190 vueltas al planeta), se demostró que los líquenes expuestos a las radiaciones cósmicas mostraron la misma capacidad fotosintética que la que tenían antes del lanzamiento, tras 72 horas de revitalización. Resultados de nuestro grupo indican que la fotosíntesis en el alga *Asterochloris erici* aislada, sometida a deshidratación progresiva, se mantiene constante hasta que ha perdido el 80 % del contenido hídrico, lo que sugiere que la gran resistencia del aparato fotosintético de este alga es debida a la implicación de un complejo mecanismo de protección, el cual posiblemente incluya componentes y/o procesos diferentes de los bien conocidos en plantas vasculares (Gasulla *et al.* , 2009).

El crecimiento y supervivencia de autótrofos fotosintéticos, como los líquenes, están limitados por la asimilación fotosintética del dióxido de carbono frente a los procesos relacionados con la respiración. Tanto la fotosíntesis como la respiración, se ven restringidos por las condiciones ambientales a las que están sometidos los líquenes, en particular, el contenido hídrico y la cantidad de luz que reciben (Green *et al.* , 2008). La concentración interna de $CO_2$ en el talo es limitante para la fotosíntesis y , debido a la baja difusión del $CO_2$ en agua, se ve disminuida por contenidos hídricos altos (Cowan *et al.* , 1992). Diferentes trabajos han estudiado la existencia de mecanismos para la concentración de carbono inorgánico (CCM) tanto de los fotobiontes aislados, como en estados liquénicos. En ellos se observó que muchos fotobiontes poseían CCMs que en el caso de cianolíquenes y líquenes con ficobiontes del género *Coccomyxa,* se basaban en la actividad de anhidrasas carbónicas (CA). En clorolíquenes con ficobiontes de los géneros *Asterochloris* y *Trebouxia*, además de detectar esta actividad, se observó que cuando eran tratadas con inhibidores de CA, aún podían realizar fotosíntesis. Además, la existencia de reducidas discriminaciones de isótopos de $^{13}C$ en estas algas, indicaba que poseían otros CCMs de forma similar a los de plantas C4 (Badger *et al.* , 1993; Palmqvist, 1993, 2000; Palmqvist *et al.* , 2002, 2008), sin embargo, no se pudo identificar la base molecular de estos mecanismos.

### 1.2.2 *Ficobiontes liquénicos, el género Trebouxia*

Las algas verdes, sobre todo los miembros de la clase Trebouxiophyceae, están frecuentemente relacionadas con eventos de simbiosis con procariotas unicelulares, plantas, animales Kerney *et al.* (2011); Leliaert *et al.* (2012) y en especial con hongos liquenizados, donde son los ficobiontes predominantes (Friedl & BÃŒdel, 2008). Especies de

los géneros *Trebouxia*, *Asterochloris*, *Trentepohlia*, *Coccomyxa* y *Dictyochloropsis* son los ficobiontes más comunes de líquenes (Friedl & BÃŒdel, 2008). Especialmente importantes son las especies del género *Trebouxia* puesto que están presentes en más del 20 % de los líquenes (DePriest, 2004). Los estudios sobre los simbiontes fotosintéticos sugiere que pueden ser importantes marcadores de las relaciones evolutivas y, por tanto, su identificación debería de ser un prerrequisito para los estudios sistemáticos de este tipo de organismos (Helms *et al.* 2001; Margulis & Barreno 2003; Peksa & Škaloud 2011).

*Trebouxia* (Trebouxiophyceae, Chlorophyta) es un género de algas verdes unicelulares, cocoides, con un gran cloroplasto central que ocupa aproximadamente un 90 % del volumen celular y un núcleo periférico. La pared celular de estos ficobiontes tiene un significado especial ya que es donde se producen las interacciones simbióticas con los micobiontes. König & Peveling (1984) encontraron que la pared de *Trebouxia* se compone de cinco capas, cada una de las cuales tiene una composición fibrilar diferente. Hay una capa interna (S1, 600nm) de celulosa mayoritariamente y alguna proteína, una capa de polisacáridos no celulósicos (S2, 160-200 nm), una capa de esporopoleninas (S3, 50-80 nm), otra capa de polisacáridos no celulósicos (S4, 40-50 nm), y una envoltura irregular exterior (S5) que contiene monosacáridos específicos de especie que deben estar relacionados con las uniones de lectina y en el reconocimiento de simbiontes. El tipo de reproducción observada es mediante aplanosporas y zoosporas (Slocum *et al.* , 1980; Ahmadjian, 1993; Takeshita, 2001). El género de microalgas *Trebouxia* se suele considerar que se reproduce de forma clonal puesto que no se han encontrado evidencias físicas (fusión de gametos, tétradas meióticas...) que corroboren la existencia de recombinación sexual. La única evidencia indirecta es el estudio de Kroken & Taylor (2000) donde encuentran una especie filogenética de *Trebouxia jamesii* con una estructura poblacional recombinante, aunque los mismos autores señalan en el texto que sus resultados podrían deberse a mutación reversa.

Los estudios clásicos basados en información fenotípica (Friedl, 1989; Ahmadjian, 1993; Takeshita, 2001) reconocen alrededor de 30 especies basándose en caracteres morfológicos y ultraestructurales (morfoespecies), pero dichos caracteres sólo son desarrollados cuando el fotobionte crece aislado del liquen en cultivos axénicos, además, estos caracteres necesitan de una tipificación o diagnóstico adecuado puesto que se ven influenciados por el medio de cultivo y las condiciones de crecimiento. En el trabajo de Peksa & Skaloud (2008) se pone de manifiesto que la morfología cloroplástica varía según etapas fisiológicas, ontogénicas o ecológicas del alga. La forma del pirenoide a nivel ultraestructural también ha sido utilizada para caracterizar las diferentes especies de *Trebouxia* (Friedl, 1989), aunque este carácter taxonómico puede llegar a ser confuso puesto que diferentes espe-

cies del género comparten la morfología pirenoidal (Kroken & Taylor, 2000). La diversidad de especies de éste género parece ser mayor a la reflejada por la morfología (especies crípticas) (Piercey-Normore, 2006; Škaloud & Peksa, 2010; del Campo *et al.* , 2010b).

En años recientes, las investigaciones ultraestructurales han sido reemplazadas por estudios moleculares. Gracias a cebadores específicos para amplificar el espaciador transcrito interno (ITS) del ARN ribosomal nuclear (ARNr) o de la sub-unidad grande del ARN ribosomal cloroplástico (ARNr 23S) (del Campo *et al.* , 2010b), permiten realizar análisis filogenéticos que, junto al aumento de secuencias de ADN de algas obtenidas de cultivos puros y extractos de ADN total liquénico (Helms *et al.* , 2001; Piercey-Normore, 2006; del Campo *et al.* , 2010b), en las bases de datos públicas, dan la posibilidad de la identificación molecular a nivel específico. Estudios filogenéticos de la región nuclear ITS son utilizados para asignar las diferentes cepas secuenciadas a partir de extractos del holobionte, a especies cuyos nombres específicos son conocidos. Aun así, uno de los grandes problemas que se presentan para la identificación de ficobiontes es que las fases sexuales son crípticas en algas liquenizadas, y por tanto la evolución de secuencias en organismos clonales dificultan la delimitación de especies Grube & Muggia (2010). Además, diferentes clados filogenéticos comparten características fenotípicas, corroborando la existencia de especies crípticas dentro de las observaciones morfológicas clásicas.

### 1.2.3 *El fotobionte Trebouxia sp. TR9 aislado del liquen Ramalina farinacea (L.) Ach.*

*Ramalina farinacea* (L.) Ach. es un liquen fruticuloso, epifito y péndulo de color verde amarillento-grisáceo, con abundantes soralios utilizados para la reproducción vegetativa. Es muy común en áreas del hemisferio Norte de ambiente templado o mediterráneo. Este contexto de diversidad ecológica sugiere una amplia plasticidad ecofisiológica de esta asociación liquénica con la que puede hacer frente a condiciones ambientales cambiantes y estresantes. La coexistencia en todos los talos analizados de *R. farinacea* de dos microalgas diferentes, *Trebouxia jamesii* UTEX 2233 y *Trebouxia* sp. TR9 (Figura 3), recolectados desde Norte América a Europa, pasando por las Islas Canarias (del Campo *et al.* , 2010b; Casano *et al.* , 2011; del Campo *et al.* , 2013), pueden contribuir a dar esta flexibilidad ecológica particular.

Utilizando dos métodos diferentes para aislar algas liquénicas (Calatayud *et al.* , 2001; Gasulla *et al.* , 2010), estudios ultraestructurales y gracias a los marcadores de la sub-unidad grande del ARN ribosomal cloroplástico (Figura 4A) desarrollados por del Campo *et al.* (2009) junto a la región ITS del ARN ribosomal nuclear (Figura 4B), se caracterizaron las dos algas que coexisten dentro del talo de *Ramalina farinacea*, *Trebouxia* sp. TR9 y *T. jamesii*. Además, en el trabajo

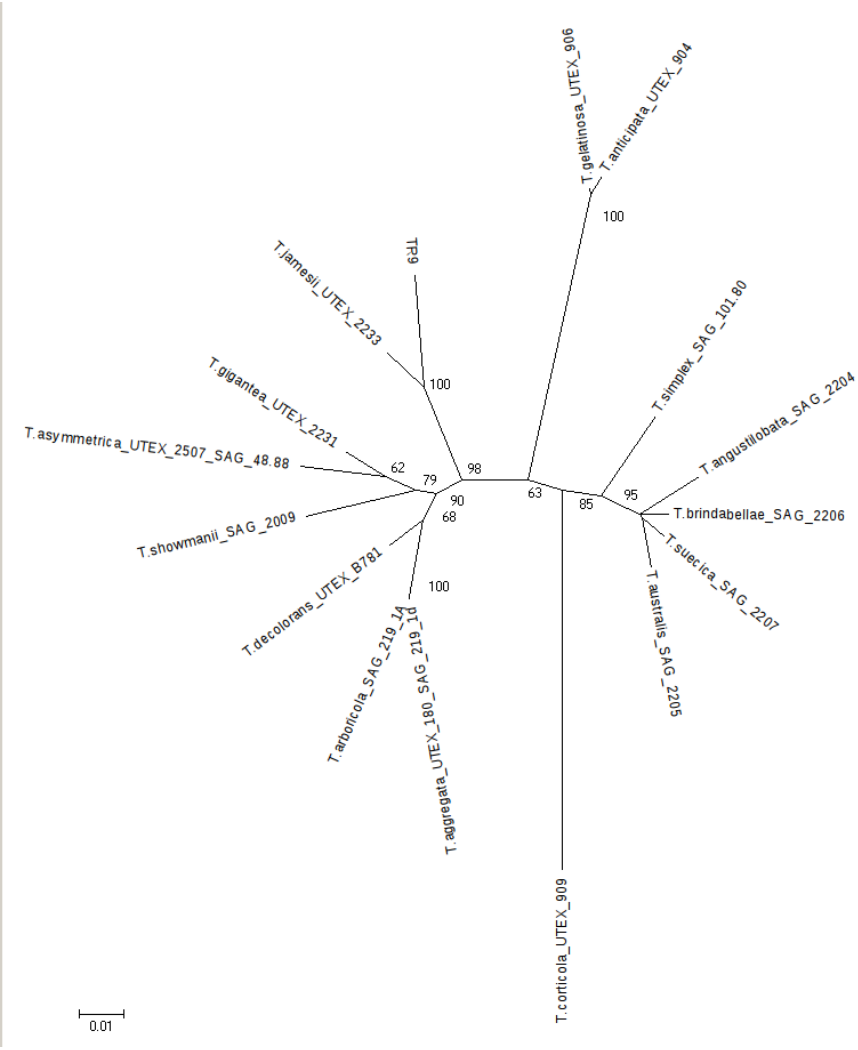



Figura 3: **Filogenia sin enraizar de las algas tipo del género *Trebouxia*.** Árbol de máxima verosimilitud del gen ribosomal cloroplástico LSU ADNr (sin intrones) y los espaciadores transcritos internos ITS1 e ITS2, junto al gen 5.8S del ADN ribosomal nuclear concatenados, bajo el modelo evolutivo GTR+I+F.

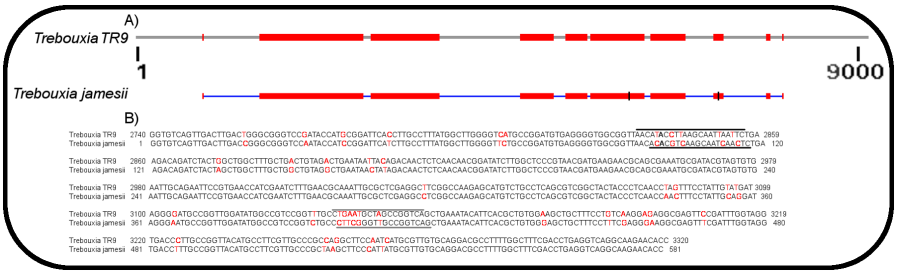



Figura 4: **Diferencias en las zonas codificantes ribosomales entre *Trebouxia jamesii* y *Trebouxia* sp. TR9. A)** Esquema del alineamiento de la sub-unidad grande del ARN ribosomal cloroplástico de *Trebouxia* sp. TR9 (este estudio) y *T. jamesii* (EU352794). En rojo se muestran las zonas alineadas y en azul las zonas presentes en *Trebouxia* sp. TR9 y ausentes en *T. jamesii*. **B)** Alineamiento de la región ITS del ARN ribosomal nuclear de *Trebouxia* sp. TR9 (este estudio) y *T. jamesii* (FJ626733). En rojo se muestran los nucleótidos diferentes y las líneas horizontales por encima del alineamiento señalan la posición de los cebadores específicos para *Trebouxia* sp. TR9 y por debajo del alineamiento para *T. jamesii*.

de Casano *et al.* (2011) se desarrollaron cebadores específicos para cada alga en la región ITS del ARN ribosomal nuclear con los que se amplificaron diferencialmente cada fotobionte a partir de ADN total del talo liquénico (Figura 4B). En el mismo estudio se caracterizaron ambas microalgas tanto al microscopio óptico como sus ultraestructuras en el microscopio electrónico de transmisión. Bajo el microscopio óptico el tamaño de *Trebouxia* sp. TR9 es siempre mayor tanto en forma liquenizada como en cultivo axénico. Bajo microscopio electrónico de transmisión la pared celular de *Trebouxia* sp. TR9 fue también mayor tanto en estado liquenizado como en cultivo, además *Trebouxia* sp. TR9 presenta una capa menos que *T. jamesii*, que presenta cuatro capas. Dentro de sus cloroplastos *Trebouxia* sp. TR9 no presenta pirenoide y contiene en su citoplasma un número mayor de vesículas electrodensas esféricas, mientras que *T. jamesii* presenta pirenoide y no posee dichas vesículas.

En el trabajo de del Campo *et al.* (2013), realizamos búsquedas de haplotipos similares a aquellos de *T. jamesii* y *Trebouxia* sp. TR9 que estuviesen presentes en otros taxones liquénicos en la base de datos del NCBI. Utilizando las secuencias del ARNr encontramos que *T. jamesii* estaba presente en los géneros *Lecanora* (*L. bicincta*, *L. cenisia*, *L. lojkaeana* y *L. rupicola*), *Ramalina* (*R. calicaris*, *R. farinacea*, *R. fraxinea*, *R. lusitanica* y *R. silicuosa*), *Tephromela* (*T. atra* y *T. grumosa*), *Amandinea* (*A. punctata*), *Anaptychia* (*A. runcinata*), *Evernia* (*E. prunastri*), *Lecidea* (*L. roseotincta*) y *Schaereria* (*S. tenebrosa*). En el caso de *Trebouxia* sp. TR9, encontramos haplotipos similares en los géneros *Ramalina* (*R. farinacea* y *R. fastigiata*) y *Tephromela* (*T. atra*).

En los estudios fisiológicos realizados con ambos fotobiontes aislados por Casano *et al.* (2011) y del Hoyo *et al.* (2011), se demostró

que cada uno de ellos tiene un comportamiento diferente frente al estrés oxidativo causado por tratamientos con Pb o con hidroperóxido de cumeno (HP-Cu, un agente productor de especies reactivas del oxígeno). *Trebouxia* sp. TR9 obtuvo un declive menor en la eficiencia máxima fotoquímica (Fv/Fm) del fotosistema II y en la disipación no fotoquímica de energía (NPQ); mayor habilidad para preservar los pigmentos clorofila a y carotenoides; mayor estabilidad para la proteína del fotosistema D1; aumento de las proteínas "heat shock 70 kDa protein" (HSP70), "glutathione reductase" (GR) y de la "superoxide dismutase" (SOD) frente a los niveles obtenidos por *T. jamesii.* Cuando se trataron a ambas especies con plomo en medio líquido (Álvarez *et al.* , 2012, 2014), ambas microalgas toleraron los tratamientos, pero cada alga ha desarrollado una estrategia diferente para ello. *Trebouxia* sp. TR9 limitaba la entrada de este metal pesado a la célula, manteniéndolo como agregados en la superficie externa de la pared celular, mientras que *T. jamesii* acumulaba mucho más plomo de forma intracelular(Álvarez *et al.* , 2012). Además, en el trabajo de Casano *et al.* (2015) se encontraron diferencias en las composiciones de las paredes celulares y en los polímeros extracelulares entre ambas especies, los cuales estaban constituidos por una proporción considerable de proteínas (cerca del 50 y del 30 % del total en *T. jamesii* y *Trebouxia* sp. TR9 respectivamente). Cuando ambas microalgas fueron expuestas a plomo, observaron un descenso en la cantidad proteica de los polímeros extra celulares en ambas microalgas y el efecto del plomo conllevó a cambios en los patrones polipeptídicos, siendo más significantes en los patrones de *Trebouxia* sp. TR9, aunque las posibles diferencias en el exoproteoma entre las microalgas liquénicas *T. jamesii* y *Trebouxia* sp. TR9 puedan estar implicadas en sus diferentes patrones de retención extracelular de metales pesados necesitan ser estudiados (Casano *et al.* , 2015).

Estudios preliminares de nuestro grupo de investigación han mostrado que la coexistencia del binomio *T. jamesii* / *Trebouxia* sp. TR9 es más frecuente de lo que pensábamos puesto que lo encontramos en líquenes de diferentes localidades de la Península Ibérica en las especies *R. fastigiata* (Catalá S. *et al.*, sin publicar) y *Tephromela atra*. Estos resultados sugieren que este binomio tiene un área de distribución que comprende -al menos- las regiones de clima Mediterráneo del hemisferio norte y Macaronésico. La detección de diferentes cepas o especies de *Trebouxia* en otros líquenes sugiere que tener clorobiontes relacionados pero fisiológicamente diferentes puede ser ventajoso para el holobionte y así poder hacer frente a los cambios ecológicos y colonizar diferentes nichos (Piercey-Normore, 2006; Muggia *et al.* , 2013; del Campo *et al.* , 2013). Por ejemplo, en el caso de *Ramalina farinacea*, la resistencia a metales pesados podría ser, en parte, debida a sus ficobiontes (Casano *et al.* , 2015).

Es muy probable que este tipo tan particular de asociación se deba, entre otros posibles factores, a los diferentes comportamientos fisiológicos de ambos ficobiontes. La secuenciación y anotación del genoma de *Trebouxia* sp. TR9 así como el análisis de su estructura y contenido de genes servirá para, por una parte, identificar rutas metabólicas y procesos fisiológicos relacionados con la eficacia adaptativa de esta especie de *Trebouxia* para modular su fisiología en respuesta a condiciones ambientales estresantes y por otra parte, permitiría estudiar aspectos evolutivos de esta especie de *Trebouxia* en relación a otras algas de la división Chlorophyta.

## 1.3 Antecedentes de genomas en algas verdes (div. Chlorophyta)

Los datos derivados de las anotaciones de los genomas secuenciados de algas verdes han proporcionado información relevante sobre muchas cuestiones fundamentales en lo referente a adaptaciones ecofisiológicas y evolutivas en este particular grupo de organismos. Estudios de genómica comparada utilizando este tipo de datos ha logrado dar respuesta a cuestiones clave sobre la evolución de las algas. A pesar de la importancia que las algas, y los ficobiontes liquénicos en particular, tienen para los análisis de procesos evolutivos, han sido poco estudiados debido al reducido número de investigadores dedicados a este tipo de organismos y a la escasez de financiación para generar este tipo de datos, al no ser considerados "organismos modelo". Sin embargo, en esta última década, se han realizado varias secuenciaciones de genomas de algas coincidiendo con el aumento del interés por las algas como fuente alternativa para biodiesel, biorremediación o para las industrias farmacéuticas, de cosméticos y de alimentación, entre otros. Los primeros genomas secuenciados fueron los del alga roja extremófila *Cyanidioschyzon merolae* Matsuzaki *et al.* (2004) y la diatomea *Thalassiosira pseudonana* (Armbrust *et al.* , 2004).

En lo referente a genomas mitocondriales de algas verdes, existen 20 genomas disponibles de los cuales tan sólo cuatro son de la clase Trebouxiophyceae: *Coccomyxa* sp. C-169 (Smith *et al.* , 2011), *Helicosporidium* sp. (Pombert & Keeling, 2010), *Prototheca wickerhamii* (Wolff *et al.* , 1994) y Trebouxiophyceae sp. MX-AZ01 (Servín-Garcidueñas & Martínez-Romero, 2012). Además, siete genomas cloroplásticos de algas verdes están disponibles en la base de datos de orgánulos del NCBI: *Chlorella variabilis* NC64A (Smith *et al*. sin publicar), *Chlorella vulgaris* (Wakasugi *et al.* , 1997), *Coccomyxa* sp. C-169 (Smith *et al.* , 2011), *Helicosporidium* sp. (De Koning & Keeling, 2006), *Leptosira terrestris* (De Cambiaire *et al.* , 2007), *Parachlorella kessleri* (Turmel *et al.* , 2009) y Trebouxiophyceae sp. MX-AZ01 (Servín-Garcidueñas & Martínez-Romero, 2012). Recientemente, Lemieux *et al.* (2014) han secuenciado 29 genomas cloroplásticos de algas clorofitas, entre las cuales se

encuentran las secuencias parciales de los genomas mitocondrial y cloroplástico de *Trebouxia agreggata,* pero éstas no permiten tener una visión global de la arquitectura de ambos genomas.

Los datos genómicos de diferentes algas de la división Chlorophyta han aumentado en los últimos años, se disponen de ocho genomas nucleares (Derelle *et al.* , 2006; Merchant *et al.* , 2007; Palenik *et al.* , 2007; Worden *et al.* , 2009; Blanc *et al.* , 2010, 2012), pero los datos genómicos referentes a las algas de la clase Trebouxiophyceae son escasos. Tan solo dos genomas nucleares, el alga endosimbionte del protozoo ciliado *Paramecium bursaria*, *Chlorella variabilis* NC64A (Blanc *et al.* , 2010) y el alga polar de vida libre *Coccomyxa subellipsoidea* C-169 (Blanc *et al.* , 2012) han sido analizados y publicados. Por otra parte, tan sólo hay un genoma del ficobionte *Asterochloris* sp., aislado del liquen americano *Cladona grayii*, y publicado en la base de datos del JGI, pero no se encuentran analizadas sus características con detalle.

En el actual mundo de la post-genómica, y en especial a partir de 2005, cuando se han introducido las técnicas de secuenciación masiva, como el número de algas genomas secuenciados es escaso en comparación con otros organismos, se hace necesario disponer de más genomas de algas para el desarrollo de información sobre la genética, la biodiversidad y otros recursos evolutivos de estos taxones (Bhattacharya *et al.* , 2015).

## 1.4 SECUENCIACIÓN GENÓMICA

### 1.4.1 *Fundamentos de la secuenciación de ADN*

El término genoma hace referencia al material genético en virus, células procarióticas y en los núcleos y orgánulos de eucarióticas. La secuenciación es el proceso de leer y descodificar los nucleótidos del ADN (o ARN) de un genoma. El proceso de secuenciación se puede dividir en tres fases: **Secuenciación**, **finalización** y **anotación**. En el proceso de secuenciación, el ADN del organismo de interés es primero aislado, luego leído por un secuenciador y finalmente se convierte en información digital que puede ser almacenada en una computadora. Una vez que los datos son producidos, se aplican diversos softwares para analizar las secuencias producidas y unir las secuencias que comparten identidad de nucleótidos entre sí para reproducir una región del genoma. Tras el ensamblaje cada región ha de ser de nuevo analizada para identificar genes, repeticiones y otras características (Mardis, 2011).

La secuenciación de ADN - sin importar el tipo de tecnología utilizada - permite la determinación de una fracción nucleotídica de una molécula de ADN a partir de grupos de secuencias cortas (referidas como **lecturas**). Cuando se quiere secuenciar un genoma completo, se suele utilizar la estrategia "Wole Genome Shotgun" (WGS) donde


Figura 5: **Analogía de libros y ensamblaje genómico.** Varias copias de cada libro (cromosoma) y carpeta (orgánulos) son rotas en tiras de papel, que tras el ensamblaje se recuperan un conjunto mayor o menor de páginas originales (Ver texto).

la molécula inicial es fragmentada en pequeños trozos de posiciones aleatorias, que a su vez, tras ser secuenciados, han de ser ensamblados entre sí para reconstruir la molécula inicial, un proceso conocido como ensamblaje **"*de novo*"**. El ensamblaje en estos caso es posible cuando el objetivo a secuenciar es sobre-muestreado y así, comparando la secuencia de una lectura con el resto de lecturas generadas es posible encontrar regiones que se superponen entre sí para unirlas de una forma progresiva hasta conseguir la secuencia de nucleótidos de la molécula inicial.

Como analogía a la secuenciación WGS podemos imaginar un libro para cada cromosoma del organismo, y en el caso de *Trebouxia* sp. TR9, una libreta para el genoma mitocondrial y otra para el cloroplástico. Cada página de los volúmenes es cortada en tiras, además hay que tener en cuenta que no todas las copias son idénticas entre sí y contienen errores de tipografía. Además, hay partes que se ven repetidas dentro de cada volumen y entre diferentes volúmenes, la tarea del ensamblado para reordenar los trozos de papel de todas las copias de cada libro se dificulta hasta el punto de, dependiendo del número de tiras obtenidas, no poder recomponer de nuevo los volúmenes originales (Figura 5).

La primera generación de secuenciadores de ADN fueron desarrollados por Caltech (Smith *et al.* , 1986) y se basaban en el método de secuenciación desarrollado por Sanger *et al.* (1977). Este método se basa en la síntesis de la cadena complementaria de una molécula de ADN molde utilizando 2'-desoxirribonucleótidos (dNTPs) y 2',3'-dideoxirribonucleótidos (ddNTPs), estos últimos marcados con un fluoróforo. Cuando se incorpora un ddNTP en la síntesis de la cadena, la DNA polimerasa deja de incorporar nucleótidos generando secuencias de diferentes longitudes (Sanger *et al.* , 1977). La proporción de dNTP/ddNTP en la reacción determina la frecuencia de terminación de cadena, y por tanto la distribución de longitudes de los fragmentos generados. Al separar por tamaños las cadenas generadas con electroforesis capilar es posible leer los nucleótidos terminales de cada fragmento con la ayuda de un detector de fluorescencia. Conforme los fragmentos generados pasan por un láser, los ddNTPs terminales

emiten en un espectro de onda diferente para cada nucleótido permitiendo su detección por un sensor de fluorescencia. La determinación del color y el orden de los fragmentos posibilita asignar la secuencia nucleotídica de la cadena de ADN molde original (Metzker, 2005).

Las tecnologías de secuenciación masiva de ácidos nucleicos de nueva generación (NGS) han marcado el inicio de una nueva era de producción de datos genómicos. Gracias a estas técnicas, la resecuenciación de organismos modelo se ha vuelto muy fácil y barata. La nueva generación de secuenciadores de ADN se diferencian de la química Sanger por la producción masiva en paralelo de secuencias de ADN, mayor rendimiento y menor coste por lectura. Son capaces de producir varios órdenes de magnitud de secuencias en comparación con la secuenciación Sanger, pero generan lecturas mucho más cortas que aportan menor información complicando el problema computacional del ensamblaje (Miller *et al.* , 2010). Existen tres plataformas que son utilizadas en el presente de forma extendida: La plataforma ROCHE/454 FLX (`http://www.454.com/`), la tecnología Illumina / Solexa Genome Analyzer (`http://www.illumina.com/`), y el sistema Applied Biosystems SOLiD® System (`https://www.lifetechnologies.com/`). Cada plataforma utiliza diferentes tipos de enzimas, de químicas para la síntesis de lecturas, de ópticas de alta resolución, de hardware y de software. Además, estas tecnologías agilizan la preparación de muestras en comparación con la secuenciación genómica Sanger basada en clones y cromosomas artificiales de bacterias (Mardis, 2008). Por contra, estas nuevas tecnologías, en comparación con la química Sanger, producen secuencias más cortas y de menor calidad media. Además existen complicaciones específicas para cada tecnología no presentes en la secuenciación Sanger como pueden ser el enriquecimiento de errores cerca del extremo 3' de las lecturas, los errores en homopolímeros (454) o sesgos en regiones ricas en GC (Illumina/Solexa) (Miller *et al.* , 2010).

### 1.4.2 *Ensamblaje de secuencias*

El ensamblaje de secuencias es el proceso por el cual, a partir de las lecturas generadas, se recompone la secuencia nucleotídica del genoma inicial. Existen dos puntos clave en este proceso: profundidad de secuencias y regiones repetidas en el genoma. Se llama **profundidad** de secuencias a la cantidad (c) de veces que un nucleótido ha sido secuenciado según la fórmula c = NL / G. Si secuenciamos un genoma de longitud G con un número N de lecturas de longitud L, cuando NL=G se obtiene 1X cobertura de secuencia. En este caso existe una baja probabilidad de muestrear todo el genoma al menos una vez. Es por ello que para ensamblajes "de novo" la cantidad de lecturas necesarias para recuperar el genoma original ha de ser alta, mayor al 10-20X.

Un aspecto crítico en el ensamblaje es la presencia de **regiones repetidas** en el genoma, y es por ello que la longitud de las lecturas es importante puesto que lecturas largas son capaces de superar dichas zonas repetidas. La presencia de repeticiones y errores de secuenciación hace que el ensamblaje de lecturas forme unidades contiguas (referidas como **"contigs**") en lugar de secuencias simples y únicas correspondientes a los diferentes cromosomas. Este fraccionamiento del genoma hace que los análisis posteriores como la predicción y la anotación de genes sea dificultosa. Por tanto, la longitud de los "contigs" depende de la cantidad de genoma secuenciado, de la longitud de las lecturas, de la estructura y la complejidad del genoma así como del número y longitud de las repeticiones presentes.

Para orientar los diferentes "contigs" entre si generados en un ensamblaje, se suele utilizar la secuenciación de extremos apareados (ROCHE 454 o Illumina "**paired-end**", PE) en la que el genoma se divide en trozos mucho mayores a la longitud de lectura - de 3 a 20 Kb - y se secuencian ambos extremos 5' y 3'. De esta forma, al conocer la distancia aproximada entre ambos extremos y la orientación de ambas lecturas se puede conocer la posición relativa y la orientación entre dos "contigs" diferentes donde cada una de las lecturas alinea, creando un puente virtual entre ellos. A la colección de "contigs" interrelacionados por el alineamiento de lecturas PE se llama **"scaffold"**. A la región entre dos "contigs" interconectados por los "mate-pair" o "paired-end" se les llama **gap** puesto que son zonas no muestreadas del genoma o partes no incluidas en el ensamblaje.

### 1.4.3 *Algoritmos de ensamblaje*

Los programas informáticos para ensamblar secuencias de los proyectos WGS conocidos como ensambladores, han de implementar algoritmos matemáticos que combinan todas las lecturas en "contigs" basándose en la similitud entre lecturas individuales. El principio básico de estos programas es que dos lecturas que se superponen entre si, presuntamente han de pertenecer a la misma región. Esta suposición es inválida para lecturas de regiones repetidas, haciendo que el problema del ensamblaje dentro de la teoría computacional sea NP-complejo (Pop *et al.* , 2002). A continuación se describen brevemente los algoritmos más utilizados para la secuenciación genómica.

#### *Algoritmos "greedy"*

Los primeros ensamblajes "*de novo*" utilizaron algoritmos **"greedy**". En estos algoritmos se computan todos los alineamientos por pares posibles entre las lecturas asignándoles un valor a cada emparejamiento potencial. El algoritmo une las lecturas combinando las que presentan el mayor valor de solapamiento. Los algoritmos "greedy" trabajan de forma local presentando el problema de que ignoran re-

laciones de larga distancia en el ensamblaje con lo que es muy difícil resolver las repeticiones (Pop *et al.* , 2002). En el ejemplo de la Figura 6, como los solapamientos de las repeticiones tienen altos valores de solapamiento, en lugar de ensamblar la secuencia original (Figura 6A), colapsa las repeticiones y los fragmentos nº1 y 3 en un "contig". La sección nº 2, intermedia a ambas repeticiones, se ensamblaría de forma errónea en otro "contig" (Figura 6B).

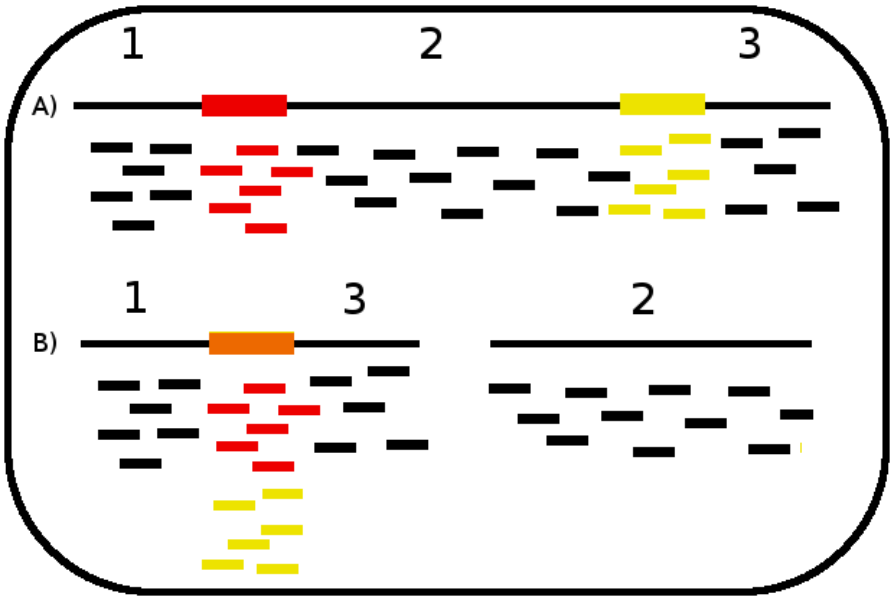



Figura 6: **Ensamblaje correcto e incorrecto de lecturas con algoritmos "greedy". A)** Ensamblaje correcto de las lecturas. Los fragmentos rojo y amarillo representan la misma repetición. **B**) Ensamblaje incorrecto de lecturas debido al algoritmo greedy. El fragmento naranja representa las lecturas colapsadas de las repeticiones.

*Algoritmos Overlap/Layout/Consensus (OLC)*

Como el campo de la secuenciación ha avanzado mucho con la aparición de las NGS, el campo del ensamblaje también ha cambiado apareciendo algoritmos basados en la teoría matemática de grafos (Earl *et al.* , 2011). Uno de estos tipos de algoritmos muy utilizados es el algoritmo de ensamblaje Overlap/Layout/Consensus (**OLC**), donde en la primera fase (overlap) se construye un grafo en el que los nodos representan las lecturas y las aristas los solapamientos entre lecturas, en la segunda fase (layout) se comprime el grafo, y finalmente en la tercera fase de ensamblaje (consensus) se determina la secuencia genómica basada en el grafo generado en los dos pasos previos (Pop *et al.* , 2002). Puesto que para generar comparaciones de solapamientos entre cada lectura y el resto es un proceso muy intenso, para aumentar la eficacia del ensamblaje estos algoritmos se basan en la noción de los "**K-mers**". Los K-mers son subsecuencias de longitud K de las lecturas de forma que, lecturas con gran similitud entre

ellas, deben compartir K-mers en sus regiones solapantes. Esta estrategia reduce el esfuerzo computacional del ensamblaje puesto que la detección de K-mers comunes entre lecturas permite acortar el número de alineamientos por pares realizado al comparar únicamente lecturas que comparten K-mers en lugar de encontrar zonas solapantes en alineaminentos de lecturas todo-contra-todo. Una desventaja de estos métodos es un descenso de la sensitividad, perdiendo verdaderos solapamientos (Miller *et al.* , 2010). Es pues que el criterio de solapamiento del OLC da como resultado dos tipos de solapamientos, los verdaderos y los falsos. Estos últimos debidos a la aparición de repeticiones en el genoma (Figura 7).

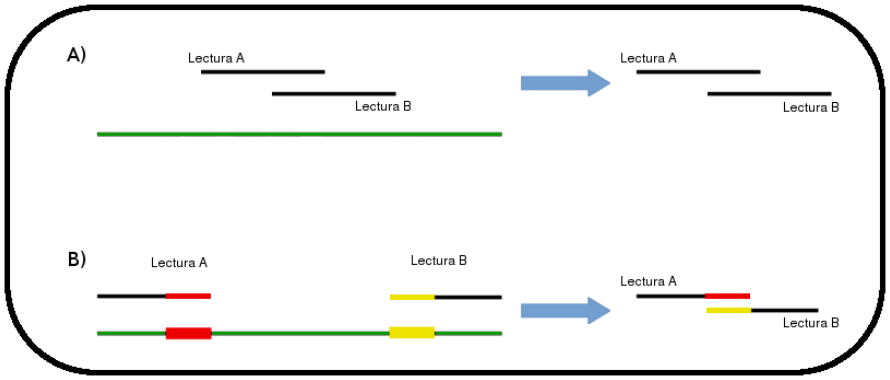



Figura 7: **Solapamientos verdaderos y falsos de los algoritmos OLC. A)** Solapamiento verdadero entre la Lectura A y la Lectura B. **B)** Solapamiento falso inducido por la existencia de una repetición compartida entre ambas lecturas, segmentos rojo y amarillo. La linea verde representa el genoma original.

Una vez construido el grafo de relaciones entre lecturas, los ensambladores han de simplificar el grafo. En un ensamblaje ideal donde el genoma no contiene ningún tipo de repetición, partiendo de cualquier nodo se podría encontrar un camino que pase por todos los nodos una única vez, lo que es conocido como un **camino Hamiltoniano**. En este hipotético caso, se lograría reconstruir el cromosoma en una sola secuencia. Con datos reales, la reducción del grafo se ve dificultada por las repeticiones y errores de secuenciación. Cuando aparecen repeticiones en el grafo, se crean bifurcaciones que conectan dos nodos no solapantes entre ellos. Al aparecer bifurcaciones, la compresión del grafo a nodos simples se para y el ensamblador comprime la repetición en un único "contig". Los "contigs" de repeticiones formados de esta forma suelen tener una alta profundidad de secuencias y están conectados a un gran número de diferentes "contigs". (Figura 8)

El la última etapa de los algoritmos OLC, tras la generación de "contigs" y/o “scaffolds”, es generada la secuencia consenso de los caminos Hamiltonianos. Comenzando en la lectura situada más a la izquierda, el algoritmo OLC computa el consenso de todas las lecturas

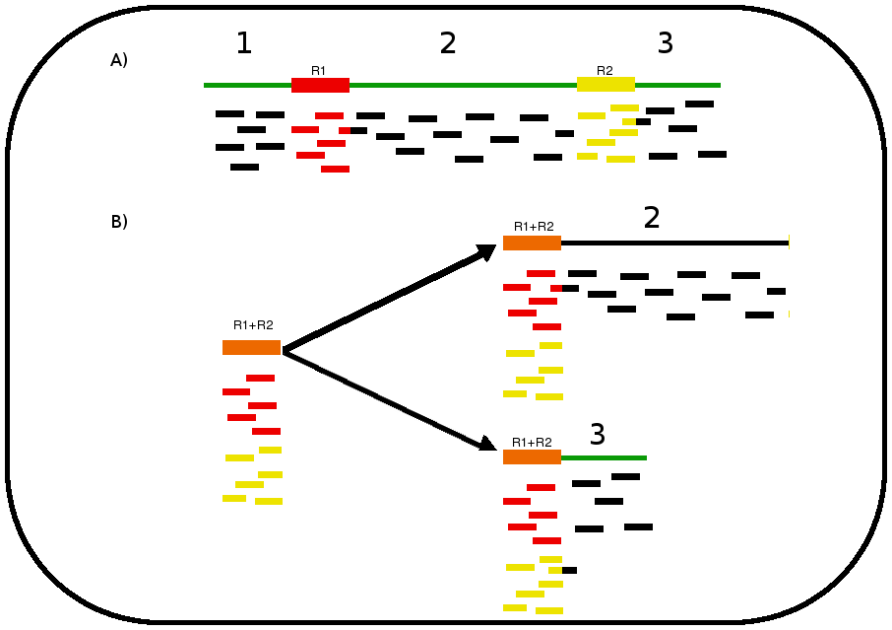



Figura 8: **Clasificación "contigs"**.

que componen cada "contig"/"scaffold". La secuencia final contendrá gaps en el genoma si existe insuficiente información de lecturas PE y/o poca profundidad de lecturas generadas.

*Algoritmos De Bruijn*

Este tercer algoritmo de ensamblaje utiliza grafos De Bruijn en lugar de grafos de solapamiento. Los ensamblajes De Bruijn también calculan "K-mers" que son usados para construir el grafo del ensamblaje. Cada "K-mer" es guardado en memoria una vez, sin importar el número de veces que ocurre en el genoma, lo que disminuye el tiempo de construcción del grafo a través del uso de **tablas "hash"**. Este tipo de estructuras de datos asocian llaves o claves con valores, en este caso, las llaves son los "K-mers" y los valores hacen referencia a las secuencias nuceotídicas de las lecturas poseedoras de dichos "K-mers". Mientras que en los algoritmos OLC hay que calcular los solapamientos entre lecturas, en los algoritmos De Bruijn no se realiza ese paso ahorrando una gran cantidad de tiempo de computación (Miller *et al.* , 2010).

En el grafo De Bruijn, las aristas representan subsecuencias únicas de longitud K dentro de las lecturas, los nodos representan subsecuencias comunes entre lecturas de longitud K-1. Por tanto, una arista conecta dos nodos si el sufijo del nodo de partida comparte con el prefijo del nodo de destino una secuencia de longitud K-2 Pop (2009).

Durante la construcción del grafo, la corrección de errores es un paso esencial de este tipo de algoritmos, puesto que son extremadamente sensibles a los errores de secuenciación presentes en las lectu-

ras ya que se generan "K-mers" no presentes en el genoma. Errores menores a K pares de bases en el final de las lecturas crean callejones sin salida o "**puntas**" en el grafo de ensamblaje (Figura 9 B). Cuando los errores se concentran en la parte central de las lecturas se forman "**burbujas**" cuando dos caminos diferentes comienzan y terminan en los mismos nodos (Figura 9 B).

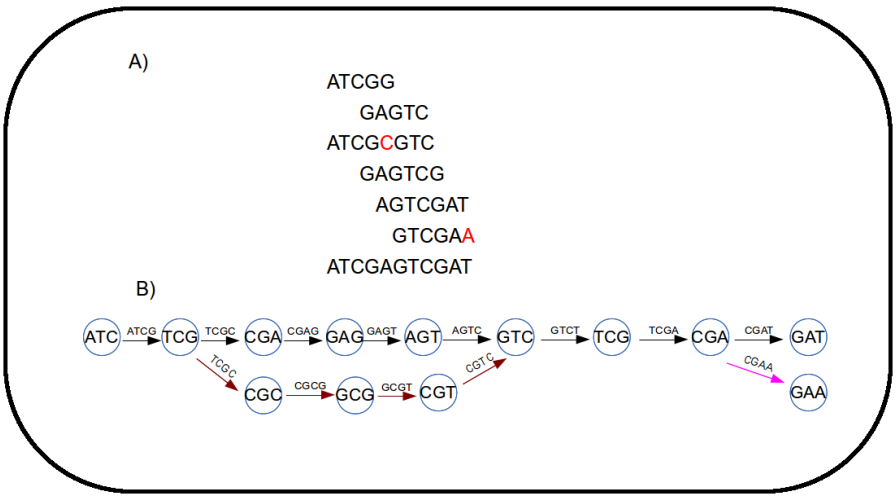



Figura 9: **Burbujas y puntas en un grafo De Bruijin.** A) Lecturas y secuencia consenso, las bases de color rojo representan los errores de secuenciación. B) Grafo De Bruijn donde K = 4. Las aristas rojas forman una burbuja y la rosa una punta.

Tras la corrección de errores, muchos algoritmos De Bruijn comprimen el grafo en "contigs", que son subgrafos no ramificados dentro del grafo. Los "contigs" se van extendiendo hasta que aparecen bifurcaciones. Las repeticiones además causan ramificaciones o diferentes caminos por los que atravesar el grafo. Los ensambladores han de identificar un **camino Euleriano**, una camino que atraviesa cada arista una sola vez. La aparición de repeticiones en el genoma crea ambigüedades en el grafo por lo que múltiples caminos Eulerianos son formados, dificultando la tarea de encontrar el camino Euleriano correcto. Un problema fundamental que aparece en este tipo de algoritmos es que al dividir las lecturas en subsecuencias de longitud K, se pierde la información de correlatividad entre "K-mers" procedentes de la misma lectura. El perder este tipo de información es un problema, puesto que la información de conexión entre "K-mer" de la misma lectura puede ayudar a recomponer el verdadero camino Euleriano ante errores de secuenciación o repeticiones. Para afrontar este problema, muchos ensambladores basados en el algoritmo De Bruijn mapean las lecturas en el grafo del ensamblaje para corregir las ramas debidas a repeticiones. Como paso final, muchos ensambladores utilizan información de las lecturas PE, de haberlas, para simplificar el grafo y construir "scaffolds".

*Conclusión*

Existen diferentes tipos de métodos de ensamblaje, con diferentes ventajas y desventajas. No existe un algoritmo perfecto, éste depende de la complejidad y del tamaño del genoma a secuenciar y del tipo de lecturas obtenidas. Los algoritmos OLC funcionan bien con lecturas largas del tipo Sanger o 454 mientras que el algoritmo De Bruijn se distingue para lecturas de secuencia corta como las lecturas de la plataforma Illumina/Solexa. Como resumen, los ensambladores para NGS - independiente del tipo de algoritmo que utilicen - comparten los siguientes puntos en común (Miller *et al.* , 2010):

1. Detección y corrección de errores de lecturas.
2. Construcción de grafos que representan las lecturas y las secuencias compartidas, ya sea en forma de solapamientos o de K-mers compartidos.
3. Reducción de caminos Hamiltonianos o Eulerianos únicos que no interconectan con otros caminos, a nodos simples en el grafo.
4. Eliminación de caminos inducidos por errores de secuenciación. Estos pueden ser bifurcaciones o burbujas.
5. Colapsan la complejidad inducida por polimorfismos.
6. Simplificación del grafo utilizando información externa al grafo. Lecturas individuales o PE actúan como restricciones en las distancias de caminos.
7. Conversión de los caminos reducidos a "contigs" y “scaffolds”.
8. Reducción de los alineamientos a secuencias consenso.

## 1.5 PIROSECUENCIACIÓN 454

En esta tesis se ha utilizado el tipo de pirosecuenciación de ADN 454, por esta razón se hace una pequeña introducción sobre ella. Este tipo de secuenciación, a diferencia de las estrategias WGS basadas en secuenciación Sanger, evita la utilización de fragmentos de ADN clonados en vectores plasmídicos (Margulies *et al.* , 2005). Para ello, tras romper la muestra de ADN y seleccionar los tamaños deseados (de 50 a 900 nt), los fragmentos son ligados a dos adaptadores (A y B) por sus extremos 3’ y 5’. Tras un paso de enriquecimiento para descartar los fragmentos con los mismos adaptadores en ambos extremos, o sin adaptadores, se capturan los fragmentos portadores de ambos adaptadores con bolas magnéticas. Estas bolas magnéticas presentan en su superficie uniones complementarias para uno de los adaptadores presentes en los fragmentos de la muestra, de forma que en cada bola se ligue un solo fragmento de ADN.

Las bolas magnéticas son utilizadas para hacer una amplificación por **PCR en emulsión** (emPCR) con una mezcla rica en grasas. De esta forma, al iniciar la reacción, en cada gota de la emulsión hay una bola magnética con un fragmento único de la muestra unido a ella. Dicho fragmento es amplificado por PCR para obtener hasta diez millones de copias. Posteriormente las cadenas de ADN se desnaturalizan, quedando las cadenas sintetizadas unidas. Cada bola es cargada en uno de los pocillos de un portaobjetos de fibra óptica. En cada uno de estos pocillos se realiza la reacción de secuenciación por síntesis basada en el protocolo de **pirosecuenciación** (Ronaghi *et al.* , 1998). En síntesis, la secuenciación de la cadena complementaria es realizada añadiendo cada nucleótido en diferentes ciclos. La liberación de pirofosfato al añadir un nucleótido a la cadena complementaria es acoplado a la producción de luz por la enzima luciferasa. Dicha luz es leída por una cámara de alta resolución CCD, monitorizando todas las bolas en tiempo real. Cuando una región homopolimérica es secuenciada, la intensidad de la luz emitida por el pocillo correspondiente es mayor a la de pocillos que han unido un único nucleótido. Cuantificando la luz emitida se puede estimar el número de bases repetidas, aunque esta pérdida de precisión de regiones homopoliméricas es uno de los puntos débiles de esta tecnología (Figura 10).

En el caso de generar **librerías MP**, los fragmentos seleccionados son de mayor tamaño y antes de añadirle los adaptadores A y B, se añade un conector para circularizar los fragmentos. Tras esto se vuelve a romper el ADN y se seleccionan fragmentos por tamaño y por la posesión del conector. Se añaden los adaptadores A y B como se expone anteriormente para continuar con el protocolo de PCR en emulsión y la pirosecuenciación. En este caso, las lecturas generadas contienen los extremos 3' y 5' de las moléculas de ADN iniciales separadas por la secuencia del conector, perdiendo longitud para cada fragmento, pero ganando la información de la distancia que separa cada extremo.

## 1.6 TECNOLOGÍA DE SECUENCIACIÓN POR SÍNTESIS ILLUMINA

Además de la pirosecuenciación 454, en el transcurso de este trabajo se ha utilizado la tecnología de secuenciación por síntesis Illumina. Esta tecnología difiere en que, tras nebulizar el ADN, se ligan unos adaptadores específicos en los extremos y tras seleccionar el tamaño idóneo (entorno a 700 nt), se ligan los fragmentos por un extremo a una placa de vidrio que contiene la secuencia complementaria de los adaptadores (Figura 10). En este caso, en lugar de realizar una PCR en emulsión, se realiza una amplificación en puente (**"Bridge PCR**"). Los fragmentos de ADN unidos por un extremo a los cebadores en una superficie sólida, se unen por su otro extremo al cebador complementario creando un "puente". Tras realizar la amplificación, se

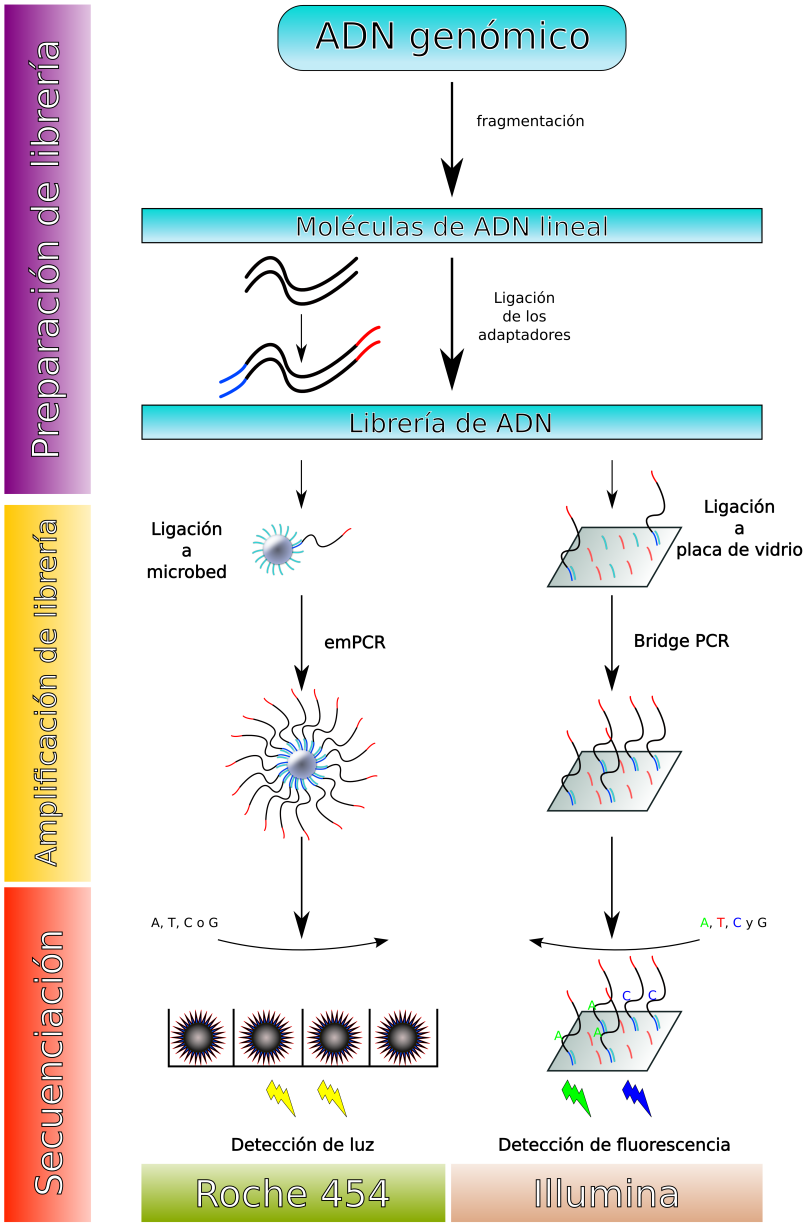



Figura 10: **Representación esquemática de la preparación de las librerías y el proceso de secuenciación ROCHE 454 e Illumina.**

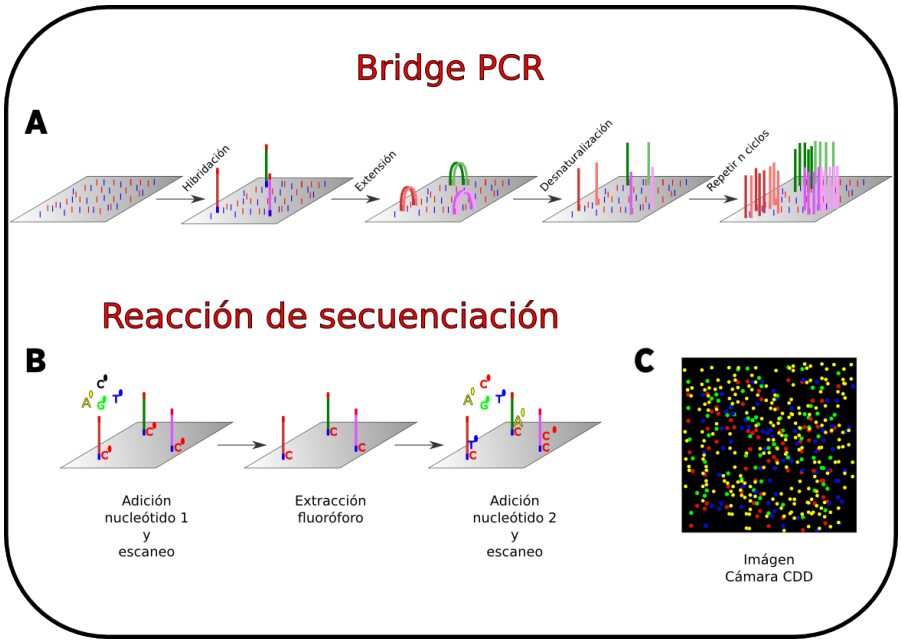



Figura 11: **Preparación de librería y reacción de secuenciación Illumina.**

crean colonias locales de ADN que serán utilizadas para una nueva reacción de amplificación, al aumentar el número de copias de cada hebra de forma local, en el siguiente paso se consigue aumentar la señal de emisión para que una cámara pueda obtener la secuencia de cada nucleótido (Figura 11 A).

Tras amplificar los diferentes fragmentos de forma local sobre la superficie de cristal, se realiza una segunda amplificación en la que se utilizan nucleótidos con un terminador reversible que emiten fluorescencias con longitudes de onda específicas para cada una de las cuatro bases. Aquellos nucleótidos complementarios a la cadena de ADN molde se añadirán, pero al tener un terminador reversible se imposibilita añadir otro nucleótido (Figura 11 B). Una cámara recoge la fluorescencia específica de cada nucleótido añadido, se elimina el terminador y se comienza un nuevo ciclo (Figura 11 C). De esta forma, todas las lecturas obtenidas son de la misma longitud y además se evita el problema de las regiones homopoliméricas de la tecnología 454, puesto que en cada ciclo se lee un solo nucleótido.

Al igual que en la secuenciación 454, en este tipo de tecnología es posible secuenciar fragmentos mayores al tamaño de lectura por ambos extremos. En este caso, se realizan librerías "paired-end" al ligar adaptadores en ambos extremos 5′ y 3′ de fragmentos de tamaño conocido. Cuando se realiza el proceso de secuenciación, tras la "bridge-PCR", primero se añade el cebador 5′ para la reacción de secuenciación, al acabar el número total de ciclos, se vuelve a secuenciar, pero esta vez desde la parte 3′ de la misma forma a cada molécula generada anteriormente. Al recibir la señal de la emisión de fluorescencia por la cámara CCD en las mismas coordenadas que la primera

reacción de secuenciación, es posible conocer la secuencia nucleotídica de cada fragmento secuenciado por ambos extremos.

# 2

# OBJETIVOS

En el liquen *Ramalina farinacea* (L.) Ach. coexisten dos especies de *Trebouxia* como ficobiontes dentro de un mismo talo (*Trebouxia jamesii* y *Trebouxia* sp. TR9). Esta forma de simbiosis tan selectiva es muy interesante como modelo de coexistencia de diferentes especies de fotoautótrofos simbiontes dentro de un único talo. Es muy probable que este tipo de asociación se deba entre otros posibles factores, a los diferentes comportamientos fisiológicos de ambos ficobiontes. En trabajos previos se ha demostrado que *Trebouxia* sp. TR9 presenta distintas respuestas inducidas por condiciones ambientales estresantes, mientras que en *T. jamesii* esas respuestas son fundamentalmente constitutivas. Todas estas características de *Trebouxia* sp. TR9 podrían ser el reflejo de una mayor capacidad para ajustar eficazmente su metabolismo y sus procesos fisiológicos a condiciones ambientales cambiantes.

Por todos estos motivos, el **objetivo principal** de la presente tesis doctoral es la secuenciación y anotación de los tres genomas (mitocondrial, cloroplástico y nuclear) de *Trebouxia* sp. TR9 utilizando diferentes técnicas de secuenciación masiva, analizar sus estructuras y contenido de genes; así como identificar rutas metabólicas y procesos fisiológicos relacionados con la eficacia de este microalga en modular sus respuestas fisiológicas frente al estrés abiótico. Además, se pretende estudiar pautas evolutivas de un alga del género *Trebouxia* en relación a las de otras algas de la división Chlorophyta.

De forma más concreta, se pueden describir los siguientes objetivos:

1. Extracción y secuenciación del ADN de *Trebouxia* sp. TR9 con las tecnologías de secuenciación masiva ROCHE 454 GS FLX Titanium, ROCHE 454 "paired-end" GS JUNIOR e Illumina Miseq "paired-end".

2. Filtrado y ensamblaje de las secuencias obtenidas por ambas tecnologías con programas bioinformáticos adecuados.

3. Determinación del tamaño y estructura general de los genomas mitocondrial, cloroplástico y nuclear de *Trebouxia* sp. TR9.

4. Identificación de genes, determinación del código genético utilizado por cada uno de los tres genomas así como la predicción de rutas metabólicas y procesos fisiológicos.

5. Comparación de los genes presentes en los genomas de *Trebouxia* sp. TR9 con otras especies de algas de la clase Trebouxiophyceae y otras clases de la división Chlorophyta.

6. Identificación y clasificación de intrones así como de posibles ORFs en algunos de los genes presentes en los genomas organulares.

7. Determinación del grado de sintenia en los genomas organulares de las algas de la clase Trebouxiophyceae, como indicador

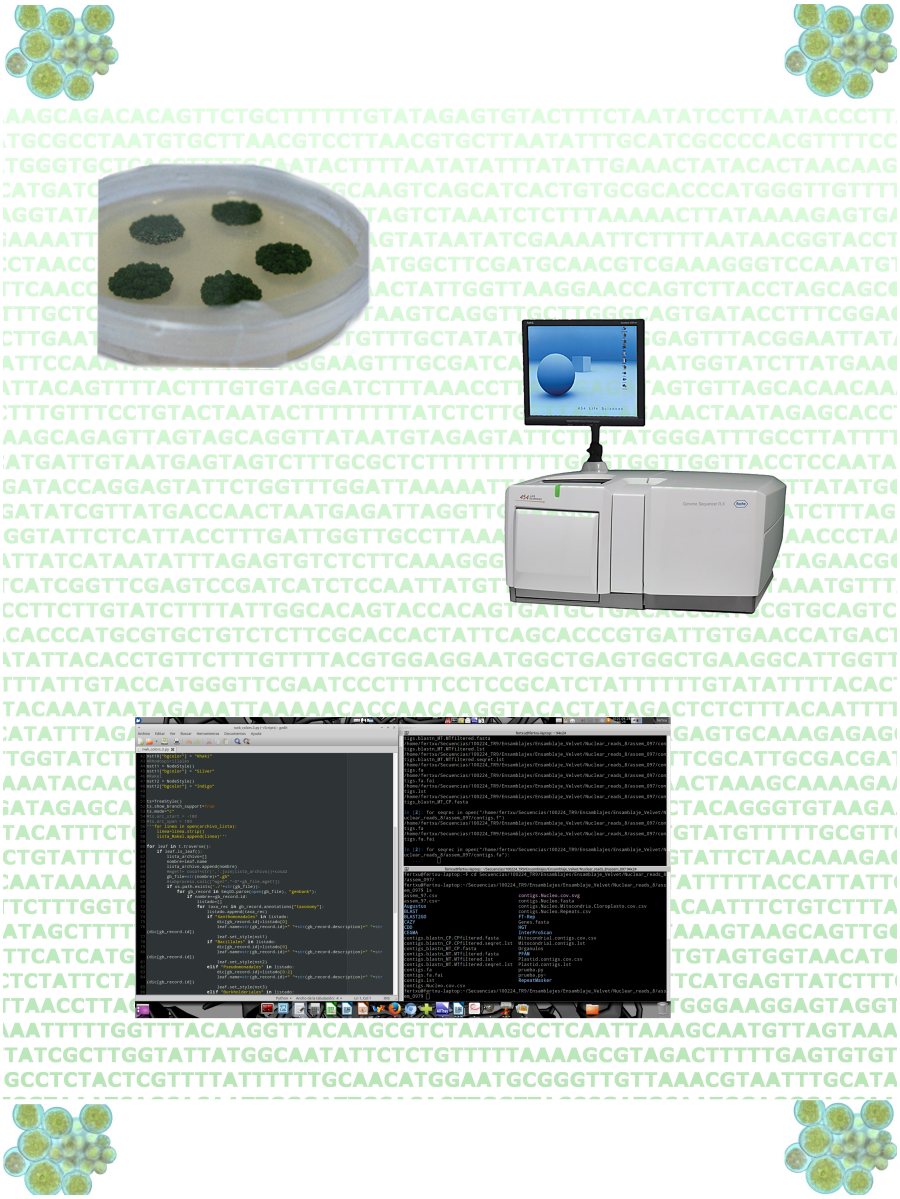
de los reordenamientos ocurridos a lo largo de su historia evolutiva.

8. Análisis filogenético basado en las secuencias de las proteínas codificadas en los genomas organulares, con el objeto de establecer la relación de *Trebouxia* sp. TR9 con otras algas de la clase Trebouxiophyceae y con las diferentes clases de la división Chlorophyta.

9. La anotación del genoma nuclear de *Trebouxia* sp. TR9 será empleada en el análisis proteómico de los polipéptidos extracelulares secuenciados de *Trebouxia* sp. TR9 y *T. jamesii* tras ser tratadas con plomo.

# 3

# METODOLOGÍA



## 3.1 MATERIAL BIOLÓGICO

Diferentes talos de *Ramalina farinacea* fueron recolectados en el Mas del Pastor en la Sierra El Toro (39° 57′ 32.34″ N - 0° 46′ 35.51″ O), Castellón, España. Los fotobiontes presentes en dichos talos fueron aislados y cultivados "in vitro" (Ver 3.2).

## 3.2 AISLAMIENTO Y CULTIVO DEL MICROALGA

Cultivos monoclonales de *Trebouxia* sp. TR9 fueron aislados siguiendo el micrométodo para aislamiento de microalgas liquénicas desarrollado en nuestro laboratorio (Gasulla *et al.* , 2010). Veinticinco mg de peso seco de *R. farinacea* fueron lavados primero con agua corriente y luego con agua destilada autoclavada. Los fragmentos fueron homogeneizados en un tubo de polypropyleno con un pistilo junto a 1 ml de tampón isotónico (0.3 M sorbitol, 50 mM HEPES pH 7.5) y se filtraron con una muselina. Se centrifugó a 490 x *g* durante 5 min.

El precipitado fue resuspendido en 200 µl de tampón isotónico y cargado en 1.5 ml de Percol® al 80 % en tampón isotónico. Tras 10 min de centrifugación a 10000 x *g* (sin freno), de las cuatro fracciones que aparecieron, se recolectaron 400 µl de la capa B, donde se encontraban las células algales casi libres de hifas fúngicas, se diluyó con la misma cantidad de agua autoclavada y se centrifugó 1000 x *g* durante 10 min. El precipitado se resuspendió en 2 ml de agua esterilizada con una gota del detergente Tween 20 y se introdujo la suspensión en un aparato de ultrasonidos a 40 KHz (Elma Transsonic Digital 470 T, 140 % ultrasound power) durante 1 min y se centrifugó 5 min a 490 x g. Este tratamiento se repitió cinco veces con el fin de eliminar contaminación bacteriana y separar las células de las microalgas. El precipitado final se resuspendió en 1 ml de agua estéril.

Finalmente, la suspensión de microalgas fue diluida 10 veces y sembrada por triple estría en placas Petri con el medio de cultivo 3xN Bold's Basal Medium (3NBBM). Las colonias algales sin contaminación se pasaron a medios de cultivo 3NBBM con glucosa (20 $gl^{-1}$) y caseína (10 $gl^{-1}$) tanto en forma líquida como en agar semisólido. Los cultivos de microalgas se mantuvieron en una cámara de cultivo a 17°C bajo ciclos de 12 horas luz / 12 horas oscuridad (intensidad lumínica 25 µmol $m^{-2}s^{-1}$) hasta que se alcanzó una densidad apropiada de biomasa para la extracción de ADN.

## 3.3 AISLAMIENTO Y PURIFICACIÓN DE ÁCIDOS NUCLEICOS

La extracción de ADN se realizó siguiendo el protocolo de Ausubel *et al.* (1989), los fotobiontes fueron molidos con nitrógeno líquido y lisados con tampón de extracción (100mM Tris–HCl pH 8.5, 10 mM NaCl, 50 mM EDTA, 2 % SDS) en un mortero. Seguidamente se reali-

zó un tratamiento con Proteinasa K (100 µg/ml) a 50ºC durante 30 min. A la solución se le añadió un volumen igual de tampón saturado fenol:chloroformo (1:1) y las fases se separaron por centrifugación. La fase acuosa fue transferida a un nuevo tubo y se añadió un volumen igual de cloroformo y se incubó en hielo durante 5 min. Tras centrifugar se recolectó la fase superior y se añadieron dos volúmenes de etanol al 70 % para precipitar el ADN por centrifugación. El precipitado se lavó con etanol al 70 %, se dejó evaporar a temperatura ambiente y se hidrató con tampón TE (10nM Tris-HCl, 1mM EDTA pH 8) durante la noche a 4ºC.

Al día siguiente se añadió ARNasa (100 µg/ml) y se incubó 20 min a 37ºC. Tras este tratamiento, SDS (0.5 %) y proteinasa K (100 µg/ml) se añadieron a la solución y se incubó 1h a 50ºC. La precipitación de proteínas se realizó añadiendo acetato sódico (3M) y un volumen igual de tampón saturado de fenol:chloroformo (1:1) durante 10 min en hielo. Tras centrifugar, la fase acuosa se transfirió a un nuevo tubo y se añadió un volumen equivalente de chloroformo y se incubó en hielo durante 5 min y se centrifugó. La fase superior fue transferida a un nuevo tubo y el ADN se precipitó con dos volúmenes de etanol al 70 % y centrifugación. El precipitado se lavó con etanol al 70 % y se dejó evaporar a temperatura ambiente, finalmente se rehidrató en tampón TE.

Además se realizaron extracciones de ADN con el Dneasy Plant Mini Kit y de ARN con el RNeasy Plant Mini Kit (Quiagen, Hilden, Germany), siguiendo las instrucciones del fabricante respectivamente.

## 3.4 DISEÑO DE CEBADORES

Para el diseño de cebadores se utilizó las aplicaciones web Oligo Calc (http://www.basic.northwestern.edu/biotools/OligoCalc.html) (Kibbe, 2007). Puesto que la secuenciación 454 presenta problemas a la hora de resolver homopolímeros, se evitaron dichas zonas. Además, siempre se intentó que la temperatura de unión del cebador a la cadena de ADN se situase, tanto para el cebador directo como para el reverso, entorno a los 58,4ºC, que el contenido en GC fuese superior al 50 % y que no se formasen horquillas de más de cuatro nucleótidos en cada cebador.

## 3.5 AMPLIFICACIÓN POR PCR DEL ADN AISLADO

Las PCRs realizadas se llevaron a cabo en los termocilcadores 96-well LabCycler (SensoQuest Biomedizicnische Elektronik) y MasterCycler personal (Eppendorf). Para la reacción se utilizó la EmeraldAmp GT PCR Master Mix (Takara Bio Inc., Shiga, Japan) que incluye un tampón optimizado, la enzima polimerasa, una mezcla de los dNTP's,

colorante de carga de gel (verde), y un reactivo de densidad en formato 2X PCR de mezcla maestra.

Las reacciones se llevaron a cabo tanto en volúmenes de 25 como de 50 µl. Las proporciones de los reactivos fueron, 1µl de ADN molde, una concentración final de 0.2µM del cebador directo y del cebador reverso, 1 µl de dimetilsulfóxido (DMSO), 8,5 µl de $H_2O$ desionizada estéril y 12,5 µl del mix de EmeraldAmp® GT PCR Master Mix (2 × Premix). En el caso de las reacciones de 50 µl, las cantidades fueron el doble de las expuestas anteriormente.

Cuadro 1: **Condiciones de PCR**

| Temperatura (ºC) | Tiempo (minutos) | Número de ciclos |
|---|---|---|
| 90 | 2:00 | 1 |
| 90 | 0:30 | 30 |
| 50 - 60 | 0:30 | 30 |
| 72 | 2:00 | 30 |
| 72 | 4:00 | 1 |

## 3.6 AMPLIFICACIÓN POR PCR EN TIEMPO REAL PARA LA ESTIMACIÓN DEL TAMAÑO NUCLEAR

Para estimar el tamaño genómico de *Trebouxia* sp. TR9 se ha utilizado la técnica basada en PCR en Tiempo Real (RT-PCR) con la fórmula tomada de Armaleo & May (2009): $LG = (QG/QP) * LP * E^{(CtG-CtP)}$. Donde QG y QP son las cantidades de ADN utilizadas en la RT-PCR tanto del ADN genómico, como del producto de PCR respectivamente, LG y LP son las longitudes de los productos de PCR finales tanto partiendo desde la extracción de ADN total, como del producto de PCR. La letra E corresponde a la eficiencia de la RT-PCR. En la Tabla 2 se muestran los diferentes oligos que han sido diseñados basándose en las secuencias nuceotídicas de los "contigs" ensamblados de *Trebouxia* sp. TR9 y que fueron testados por si aparecían productos secundarios en las reacciones de RT-PCR.

La amplificación del fragmento de partida se realizó en un volumen total de 50µl que contenían: 3µl de ADN genómico de *Trebouxia* sp. TR9, 3µl de cada oligo (10µM), 16µl de $ddH_2O$ y 25µl de Taq EmeraldAmp® Master Mix (Shiga, Japón). Las condiciones de reacción fueron: un ciclo a 90ºC durante 2 min; 40 ciclos de: 90ºC durante 30s, 58.4ºC durante 60s y 72ºC durante 2min. Estos ciclos fueron seguidos de una extensión final de 4min a 72ºC. Los productos obtenidos fueron visualizados en un gel de agarosa al 1 % y se purificaron con el QIAquick PCR Purification Kit (QUIAGEN) siguiendo las instrucciones del fabricante.

Cuadro 2: **Parejas de cebadores utilizados para realizar la estimación del tamaño genómico de *Trebouxia* sp. TR9.**

| Oligo Directo | Secuencia | Oligo Reverso | Secuencia |
|---|---|---|---|
| TR9_AhpC_F | GGAACTCCTGGCTGCACA | TR9_AhpC_R | TGATACCAAAGCTCTTCCTCAA |
| TR9_Catalase_F1 | CATGACATCTCTCACCTGAC | TR9_Catalase_R2 | AACCAGATCCCAGTTGCCTT |
| TR9_Catalase_F2 | GTGATGACAAGATGCTGCAG | TR9_Catalase_R3 | GAAGAAGTCAGCAATTCGCC |
| TR9_Catalase_F2 | GTGATGACAAGATGCTGCAG | TR9_Catalase_R4 | CATCTCTGTGCATGAAGTTCAT |
| EF-2_F | GCTGGTCTTCACGGTTGG | EF-2_R | CCCCACAGCTTCTCCATC |
| EF-6_F | GATCTTGATGAGCTATCATCAC | EF-6_R | CTGCCCAGTCATTTGCAACA |
| FER1_F | GGTTAAGGATGAGCTGTCAG | FER1_R1 | GCTCATTGATTGCATATTCTGC |
| FER1_F | GGTTAAGGATGAGCTGTCAG | FER1_R2 | CTGGAAGTCCATCAAGGTCT |
| HSP90A_F1 | AGGTATCACAGCACCAAGTC | HSP90A_R1 | CTTCTCAATGAAGGGGGAGT |
| ACT_F | GAGCGAGGTTACAGTTTCAC | ACT_R | GGTTGAACTTGAAGCAGCAG |
| ACT_F1 | GTGCCCATCTATGAGGGTTA | ACT_R | GGTTGAACTTGAAGCAGCAG |
| PEPC_F1 | ATGTACCAGAGCTGGCCATT | PEPC_R1 | TGCATAGCTGCTGAAATCCC |
| PEPC_F1 | ATGTACCAGAGCTGGCCATT | PEPC_R2 | GCCAAGCGCTTGCTGCAT |
| PEPCK_F1 | GTCTTGCCTGCTTTGCATAC | PEPCK_R1 | CCCAAAGAAGAGGGTGACAT |
| PEPCK_F2 | TGGTCATCCTTGGTACTCAG | PPCK_R2 | TATATTGGAGATTGCTTATTCGTC |
| TUB2_F1 | GAGGTTGATGAGCAGATGCT | TUB2_R1 | GGGAGGGATGTCACACAC |
| TUB2_F2 | GGTGAGGGCATGGATGAG | TUB2_R2 | CTTCCTCCTCATCAAACTCG |

Para todas las reacciones de RT-PCR se han utilizado cuatro concentraciones diferentes ($10^{-2}$,$10^{-3}$,$10^{-4}$,$10^{-5}$) para cada gen tanto para los fragmentos amplificados por PCR , como en las reacciones con ADN genómico. Todos los análisis de las muestras han sido llevadas a cabo con tres réplicas para reducir la variación generada en las reacciones de PCR. Cada mezcla de 20µl en total de la mezcla de RT-PCR contenía 0.4 µl de cada oligo (10µM), 0.4µl de ROX Reference Dye, 2 µl de molde de ADN, 6.8 µl de $ddH_2O$ y 10 µl de SYBR Premix Ex Taq. Las condiciones de la reacción de RT-PCR fueron: 30s a 95ºC seguidos de 40 ciclos de 5s a 95ºC y 30s a 60ºC en un ABI StepOnePlus™ real time PCR instrument (Applied Biosystem Inc, USA). Las eficiencias de la reacción se calcularon utilizando la pendiente de la curva lineal de los valores de Ct para el logaritmo de las series de diluciones en todas las RT-PCR. Aquellas parejas de oligos que presentaron un $R^2$ mayor al 0.99 y que no contenían productos no específicos, fueron utilizadas para el cálculo de la estimación del tamaño genómico de *Trebouxia* sp. TR9.

El número de copias del gen rRNA se determinó mediante PCR en Tiempo Real siguiendo la fórmula de $2^{(\text{Ctref} - \text{Cttest})}$ (Wang *et al.* , 2011). El gen FER1 fue elegido como gen de referencia y su Ct se definió

como Ctref. El fragmento ITS1, que se localiza en cada gen del ARN ribosomal, se consideró como gen de ensayo y su Ct se definió como Cttest.

## 3.7 SEPARACIÓN ELECTROFORÉTICA Y PURIFICACIÓN DE PRODUCTOS DE AMPLIFICACIÓN

Los productos de amplificación por PCR obtenidos se visualizaron en geles de agarosa al 0,8 - 1 %, según los fragmentos fuesen de mayor o menor tamaño, respectivamente. Cuando aparecieron productos no específicos, las bandas correspondientes al tamaño esperado, fueron cortadas con una cuchilla del gel de agarosa. El ADN fue purificado utilizando los kits Illustra GFX PCR DNA cuando no se necesitó cortar ninguna banda y el kit Gel band Purification kit (GE Heathlcare Life Science, Buckinghamshire, England) en el caso de que apareciesen diversas bandas.

## 3.8 SECUENCIACIÓN DEL ADN AMPLIFICADO Y PURIFICADO

El ADN obtenido para la pirosecuenciación se envió a secuenciar 1/4 de placa bajo la plataforma de secuenciación ROCHE 454 GS FLX Titanium sequencer en LifeSequencing S. L. en el parque científico de la Universidad de Valencia en 2010. La misma extracción de ADN fue utilizada en 2012 para realizar una secuenciación "paired-end" con un tamaño medio de inserto de 3 Kb bajo la plataforma de secuenciación ROCHE 454 GS JUNIOR en la Unidad de Genómica SCSIE-Universitat de València. Una segunda extracción de ADN de *Trebouxia* sp. TR9 fue secuenciada en 2014 bajo la plataforma de secuenciación Illumina MiSeq con una librería "paired-end" que constaba de un tamaño de inserto medio de 1 Kb en el Parque Científico de Madrid.

Las PCRs realizadas fueron secuenciadas en un ABI 3100 Genetic analyzer utilizando kit ABI BigDyeTM Terminator Cycle Sequencing Ready Reaction (Applied Biosystems, Foster City, California). Las secuencias obtenidas fueron procesadas para eliminar zonas de baja calidad con los programas preagap4 y gap4 del paquete bioinformático Staden (Staden *et al.* , 1999).

## 3.9 ENSAMBLAJE DE LECTURAS Y ANÁLISIS DE "CONTIGS"/"SCAFFOLDS"

### 3.9.1 *Ensamblajes "de novo"*

Lecturas ROCHE 454 GS FLX Titanium (2010)

Estas lecturas fueron ensambladas con dos programas diferentes: GS De Novo Assembler v 2.5 y MIRA V3.2.0 (production version) (Chevreux *et al.* , 2004) para comparar los resultados obtenidos por cada ensamblador y seleccionar el mejor ensamblaje para posteriores

análisis. En el primer caso, al ser un programa nativo para la secuenciación ROCHE 454, se utilizó directamente como entrada para el ensamblador el archivo sff obtenido por el centro de secuenciación. En el caso del ensamblador MIRA, primero se extrajeron las lecturas obtenidas del archivo sff y fueron procesadas para quitar los adaptadores de secuenciación con el script de Blanca y Chevreux sff_extract (http://bioinf.comav.upv.es/sff_extract/). Puesto que las secuencias contenían junto al adaptador A el MID de secuenciación nº8 la opción [*min_left_clip=15*] fue utilizada para enmascarar dicha secuencia. Una vez obtenidas las lecturas sin los adaptadores, se realizaron tres tipos de ensamblajes diferentes con el ensamblador MIRA V3.2.0, cada uno con las opciones "accurate", "normal" y "draft" respectivamente. Se realizó además otro ensamblaje con este programa, en este caso las lecturas con los extremos 5' enmascarados fueron filtradas por calidad y se recortaron los extremos 3' de baja calidad con el programa LUCY v1.20p (Chou & Holmes, 2001) con las opciones por defecto. Estas lecturas fueron ensambladas con la opción "accurate".

Tras comparar los resultados de los cinco ensamblajes diferentes, se seleccionó, basándonos en el estadístico N50, entre otros, el ensamblaje realizado con el programa MIRA con las lecturas filtradas por calidad con el programa LUCY para análisis posteriores y el ensamblaje del genoma mitocondrial (Ver 3.9.2).

Lecturas ROCHE 454 "paired-end" GS JUNIOR (2012)

Las lecturas "paired-end" obtenidas en 2012 en formato sff fueron ensamblada conjuntamente con las obtenidas en 2010 con el ensamblador GS De Novo Assembler v2.6 puesto que este ensamblador toma la información "paired-end" para crear "scaffolds" de "contigs" interconectados. Se probaron diferentes ensamblajes para los valores 80, 85, 90, 95 y 100 en la opción [*overlapMinMatchIdentity*]. El valor 20 para la opción [*minimumReadLength*] fue seleccionado ya que al quitar el conector de enlace entre la parte 5' y 3' de las lecturas PE, la parte 3', al estar al final de la secuencia, perdía longitud útil debido a la baja calidad de secuencia al final de la lectura.

Lecturas Illumina Miseq "paired-end" (2014)

Las lecturas originadas bajo la plataforma Illumina MiSeq se vieron sujetas a un filtrado por calidad con el software MIRA4, en esta ocasión se realizó tan solo el primer pase de limpieza de lecturas y con el módulo MIRA_convert se extrajeron las lecturas procesadas del archivo en formato caf que se obtuvo durante este proceso. Tan solo las lecturas mayores a 20 nt y que ambos extremos pasaron el filtro de calidad, fueron utilizadas para los subsiguientes ensamblajes. Se realizaron diferentes ensamblaje "de novo" con el software Velvet (Zerbino & Birney, 2008) utilizando las lecturas Illumina filtradas y como guías los "scaffolds" -sin los pertenecientes al cloroplasto o a la mitocondria- y los "contigs" que no formaban "scaffolds" del ensamblaje "de novo" de lecturas 454 realizado en 2012. Para comprobar

Cuadro 3: **Cebadores mitocondriales diseñados para unir "contigs" y comprobar circularidad del genoma.**

| Oligo Directo | Secuencia | Oligo Reverso | Secuencia |
|---|---|---|---|
| MT_61K_769_F | AGTTTACGGAATTATAACAGCG | MT_61K_1838_R | TACGTTGATTTAGCAAACCAATG |
| MT_61K_23618_F | AGTAGAGACACAACATCATTAAC | MT_61K_24958_R | GAGCTGACGACAGCCATG |
| MT_61K_59310_F | TGTGTTTACCTATTTCACCAAG | MT_61K_60869_R | GAAAGTGGCTCTTCCAGCA |
| MT_10K_431_F | ACACCTAGTTGGTATTGCTTTG | MT_10K_654_R | GGTGTTTGAAAGATAGACTGCA |
| MT_10K_9453 | GCATATCGTCAAATGTCATTG | MT_10K_9858_R | CAAGTATTGAGTAGCGGCGT |
| MT_10K_654_R | GGTGTTTGAAAGATAGACTGCA | MT_61K_1838_R | TACGTTGATTTAGCAAACCAATG |
| MT_10K_654_R | GGTGTTTGAAAGATAGACTGCA | MT_61K_59310_F | TGTGTTTACCTATTTCACCAAG |

qué tamaño de palabra (K-mer) era el más correcto para generar el grafo de Bruijn y el posterior ensamblaje, se probaron los "k-mer" impares del 21 al 133.

### 3.9.2 *Ensamblaje mitocondrial*

*Secuenciación 2010*

Tras los ensamblajes "de novo" realizados en 2010, se seleccionó el ensamblaje realizado por MIRA con las lecturas filtradas con el programa LUCY puesto que el estadístico N50 y el "contig" más largo fueron los mayores de todos los ensamblajes. Los "contigs" mitocondriales fueron seleccionados utilizando el algoritmo BLASTn, BLASTx y tBLASTx (Altschul *et al.* , 1997) contra una base de datos local de los genomas y proteínas mitocondriales de Viridiplantae descargados del NCBI. Se obtuvieron dos "contigs" pertenecientes a la mitocondria de *Trebouxia* sp. TR9. Para realizar las uniones entre los diferentes "contigs" y corroborar la circularidad del genoma, se diseñaron cebadores específicos para cada "contig" en sentido directo en los extremos 3' y en sentido reverso en los extremos 5' (Tabla 3), se probaron las ocho posibles combinaciones de orientación entre los "contigs" y finalmente, la combinación TR9_MT_61k_59310_F - TR9_MT_10k_654_R unió ambos "contigs". El resto de cebadores (Tabla 3) se utilizaron para corroborar la circularidad del genoma.

*Secuenciación 2012*

La secuencia mitocondrial obtenida con las primeras lecturas 454 secuenciadas en 2010 fue utilizada como molde para mapear las nuevas lecturas 454 "paired-end" con el mapeador GS Reference Mapper v2.6 y así, puesto que la molécula de ADN mitocondrial es mucho mayor que el tamaño del inserto de las lecturas paired-end, se estimó el tamaño medio real de inserto de la librería.

### 3.9.3 *Ensamblaje cloroplástico*

*Secuenciación 2012*

Gracias a la información de distancias de las lecturas paired-end, tras el ensamblaje "de novo" con el programa GS De Novo Assembler v2.6, se filtraron los "contigs"/"scaffolds" del genoma cloroplástico utilizando el algoritmo BLASTn, BLASTx y tBLASTx (Altschul *et al.* , 1997) contra una base de datos local de los genomas y proteínas cloroplásticas de Viridiplantae descargados del NCBI. Para observar las interconexiones entre "scaffolds" y determinar las uniones entre "contigs", se utilizó el script bb.454contignet (Iorizzo *et al.* , 2012) [http://www.vcru.wisc.edu/simonlab/sdata/software/#contignet] .

Puesto que el algoritmo de ensamblaje utilizado por el gsAssembler se basa en la unión de lecturas por medio de dos rondas del algoritmo Overlap Layout Consensus (OLC), muchos de los posibles problemas locales de ensamblaje durante la segunda ronda del OLC (repeticiones y errores de secuenciación), pueden producir la división de lecturas y disponer cada parte de la lectura en diferentes "contigs", creando así "contigs" más pequeños fragmentando el ensamblaje. Para solucionar esto, todos los "contigs" de los "scaffolds" pertenecientes a cloroplasto junto a los "contigs" adicionales recuperados por el script bb.454contignet fueron mapeados con el ensamblador MIRA V3.2.0. Este ensamblador filtra las regiones de superposición entre las lecturas utilizando un algoritmo de alineamiento Smith-Waterman bandeado y por tanto, cuando en el final de los "contigs" generados por gsAssembler existen pequeñas repeticiones que son posibles de sobrepasar con el tamaño de lectura, MIRA no rompe la lectura como en el caso de gsAssembler, y por tanto alarga el "contig".

Tras el mapeo, cuando fue posible, se unieron "contigs" adyacentes en los que sus extremos se superponían gracias a la información de las lecturas añadidas por MIRA y las interconexiones entre los "contigs" de los "scaffolds" obtenidas con el script bb.454contignet. Para ello se utilizó la opción "*find internal joints*" del programa gap4 del paquete Staden (Staden *et al.* , 1999).

En los casos donde debido a errores en el extremo 3' de las lecturas no se pudo unir dos "contigs" directamente, se diseñaron cebadores en cada extremo de los "contigs" (Tabla 4) y se amplificaron por PCR. Las secuencias obtenidas por secuenciación Sanger se utilizaron como guías para unir dichos "contigs" dentro del programa gap4. Además se tuvo en cuenta la profundidad de lectura de cada contig para poder discernir las repeticiones colapsadas en un solo contig y poder separarlas con ayuda de PCRs.

Cuadro 4: **Cebadores cloroplásticos diseñados para unir "contigs" y comprobar circularidad del genoma**

| Oligo Directo | Secuencia | Oligo Reverso | Secuencia |
|---|---|---|---|
| rbcl_F2 | TCGTAGTCGTGGTGTTTACTTT | CP_1_R | CATTGTAACGAACTGTACTAT |
| CP_1_F | GCTTGCTTACTTTACTTAGTTTAG | CP_2_R | AACAGTCAATGGTTCCCCTG |
| CP_1_F2 | TGCTTGTCCGCAGTTTTGTTA | CP_2_R | AACAGTCAATGGTTCCCCTG |
| CP_2_F | CAGGTTTGTTAGGATGACCAT | CP_3_R | TAGATAAGTGGTTGTTTGTGTTC |
| CP_4_F | GTATAGCAACAACTTGGTCCT | CP_5_R | CATTGTAACGAACTGTACTATTC |
| CP_5-6_F | GGATTGCTTTCATAACATTGCC | CP_5-6_R | GCTACGCATCAGTCAGCTC |
| CP_7_F | CTTAGCCGGATCAACCTCAT | CP_8_R | CCTGGACAAACCCTTGTTCT |
| CP_7_F2 | CCTTTGCAGTTAAAATATCTTATAAG | CP_8_R | CCTGGACAAACCCTTGTTCT |
| CP_8_F | CTTCTTGTCCTTTACTTGGAAC | CP_9_R | GATATTTTTGTTGTATTGCCAGCT |
| CP_9_F | GTTGTTTGGATTTGTTATCTTGGT | CP_11_R | ACAGATAAGCCAATTCTCTATAAG |
| CP_10_F | TTGCTTTAGCGCATCCTACTAA | CP_11_R | ACAGATAAGCCAATTCTCTATAAG |
| CP_12_F | CTAGTATACCAGTTCTCGTTC | CP_13_R | ATAAACCTTGTGCCACGCAAA |
| CP_12_F2 | GATATACTAACAGACCGATTGAAT | CP_13_R | ATAAACCTTGTGCCACGCAAA |
| CP_12_F3 | TTTGCTTGGCTAACCAAGGG | CP_13_R2 | CTTTTTTTTTGGTTTTTTTACTTTTTTTTTT |
| CP_14_F | GCTCTACTATGTATCTGATGC | CP_15_R | GTTGTTTTACTTCTTTGCCATGC |
| CP_24_F | TATCAGCTTCCAACCCATAAAG | CP_25_R | GGTACCGTCCTTTCTTTTGG |
| CP_25_F | TGCAAAGGGAAGTCAATAAAACAA | CP_26_R | CTTGTCCTTCTTTGCCGTG |
| CP_33_F | GGACAAAGAAGTAAAACAACAATC | CP_34_R | TGCTTCTTTTGTTAGTTGTTTGAC |
| CP_34_F | GTATGGATTAGTTGATCAAGTAG | CP_37_R | GGGGCGCTTTGATTGTTGTT |
| CP_38_F | CCTGACAAACCTTGTGCTAC | CP_39_R | AAAAGAAGCACTGACTGATGC |
| CP_39_F | GCATCAGTCAGTGCTTCTTTT | CP_40_R | AAACGGCTGATAATAACCAACTT |
| CP_40_F | GCAAAGCTCCACCCTCG | CP_41_R | GAGACACAGTCAGTGGTTG |
| CP_54_F | GCAAGGTTTGTATAAGATTTTAACA | CP_56_R | GGGACAAAGAAAAGATAAATTCAC |
| CP_60_F | GAACAATGCAAACAAGAAGGG | CP_62_R | GCAATCAAGAACAATTATGCGC |
| CP_62_F | GTGTTTCTTTGCCATGCCTG | CP_63_R | GTTTGTCAAGGTTTGTCAGG |
| CP_64_F | GGGAGACGAACTTTGACAAG | CP_65_R | TTTGCACGGAAAATTAGATATGTTT |
| CP_67_F | GTCCTTTGCAGTTAAAATCATAAC | CP_69_R | GATGCGCAGCACTTATCATAG |
| CP_69_F | GTTCCCCTTCTCCGAGAA | CP_70_R | AGCCCTATGCTTTTTGTGTAAG |
| CP_70_F | GTCTGAAACAAACCTTTAACAAAC | CP_71_R | CCACGAACTTTGTCAAGATTTTA |
| CP_71_F | CCCAGACCTTGTGGTACG | CP_72_R | ATCCATGCCTGATGCCCTA |
| CP_72_F | CCATTGGTTAAAAGCGCATG | CP_73_R | ATCACAAACAGAGAAGGGATAC |
| CP_73_F | ATGCTGACCCTTTGCCATTG | CP_74_R | GGCTTAGAACGTTAAAATCATAGA |
| CP_76_F | GGTTTGTATAAGATTTTAACTGCG | CP_77_R | GCGCATCACAAACTGCATTTG |
| CP_81_F | ACCCCCTTTGCTCTGCAC | CP_83_R | GAAAACGAAAAGAAAACCCCTTT |
| CP_90_F | TTTGTTTTTGTCAGTACTGCGC | CP_91_R | CCCGTTTCCCCTTTCCCT |
| CP_91_F | GTAGTAAATAAACATTGTGCTTTGTA | CP_92_R | TCCGATTTTAACTGCAAAGGAC |
| CP_92_F | ACCTTGAACCAAACCTATGGTT | CP_93_R | ACTGATTTGGGTATAGGGATTT |
| CP_94_F | CAGCGCATGCATGCGAAAA | CP_95_R | CTTAGTTATACTTGTTTACTTTTATTTG |
| CP_95_F | GTGGTATAAGCATGTCTAAAGAAT | rbcl_F2 | TCGTAGTCGTGGTGTTTACTTT |
| CP_278_F | CTGAAACACATTGGTGATAAAAC | CP_97_R | GTAGCACAAGGTTTGTCAATATTT |
| rbcl_R | CACCTTCTAATTTACCTAC | CP_97_R | GTAGCACAAGGTTTGTCAATATTT |
| CP_97_F | CACAATTGAAAGACCATTTAGAAAC | CP_98_R | GTCATACCTTAACTTTCTAAAACC |
| CP_98_F | GTAACAGCGCATGCATGCC | CP_99_R | GTTGCGTCTACCAATTCCG |
| CP_101_F | CAAATTGATAAAGAAACTTTAGTTAATC | CP_102_R | ACTGAAACGGTCAGTGGTTC |
| CP_107_F | CAGTGCCGTGCGAAGCAAA | CP_109_R | CCTTAAGCCCCATGCCTG |
| CP_109_F | ACCGCAGGTTTGTTCAGGTT | CP_110_R | GGAAGATTTTGTTTTATTATTTGTATC |
| CP_110_F | GATACAGTCAGGCAAAAGGG | CP_111_R | CAAACCTTAAACAAACCGTAGG |
| CP_111_F | CTGCGGACAAGAAGGCTG | CP_112_R | GCAAAGACACATGATTTACAAGG |
| CP_112_F | GTTTATCCAGGTTTGATTCAAG | CP_278_F | CTGAAACACATTGGTGATAAAAC |
| CP_112_F | GTTTATCCAGGTTTGATTCAAG | rbcl_R | CACCTTCTAATTTACCTAC |

Cuadro 5: **Cebadores diseñados para amplificar por Transcripción Inversa y secuenciación Sanger del transcrito del gen mitocondrial *rrnL*.**

| Oligo Directo | Secuencia | Oligo Reverso | Secuencia |
|---|---|---|---|
| rrnL_F1 | TGGGTGGATGCCTAGGCA | rrnL_R1 | GTCATCCCCGCCTATTGC |
| rrnL_F2 | ATAATGGGTCAGCGAGTAAATC | rrnL_R2 | GACCAGTGAGCTATTACGCT |
| rrnL_F3 | GCAGACAGACTTTGGGCG | rrnL_R3 | CCACACGGTGCTCCCG |
| rrnL_F4 | CAGTGGTTGGTGGGTAGTTT | rrnL_R4 | GTATTTTAGTACGAGTAGGCT |

## 3.10 ANOTACIÓN DE GENOMAS ORGANULARES

Para la anotación de los genomas mitocondrial y cloroplástico se crearon dos bases de datos con los genomas mitocondriales y cloroplásticos de Viridiplantae descargados de la base de datos pública del NCBI. Cada orgánulo fue alineado con el algoritmo de alineamiento local BLAST tanto de nucleótidos contra nucleótidos (BLASTn y tBLASTx) como de nucleótidos frente a proteínas (BLASTx). Los archivos obtenidos fueron importados a la herramienta de anotación Artemis 15.1.1, donde además se obtuvieron las posibles pautas de lectura abierta (ORFs). Los sitios codificantes para genes ribosomales (rRNA y tRNAs) fueron anotados con la ayuda del servidor web Rfam basándonos en la conservación de su estructura primaria y secundaria. Para dibujar la visualización del genoma mitocondrial y cloroplástico se utilizó la herramienta web GenomeVX (Conant & Wolfe, 2008), como entrada se introdujo el archivo en formato GenBank generado tras la anotación de cada orgánulo.

En el caso de la anotación del genoma mitocondrial, ha sido necesaria la realización de diferentes reacciones de transcripción reversa. En particular, se secuenció por partes el transcrito completo para el gen de la sub-unidad grande 23S ribosomal (*rrnL*) debido a que en él se presentaron diferentes intrones. La reacción de retrotranscripción se llevó a cabo con el Kit PrimeScript$^{TM}$ RT reagent Kit de la casa TaKaRa. Cada una de las reacciones se llevaron a cabo con 0,5 µl de cada uno de las parejas de oligos listados en la Tabla 5, 0,5 µl de mezcla enzimática, 2 µl de Tampón 5x, 2 µl de ARN y 4,5 µl de agua doble destilada. El programa que se utilizó fue de 30 minutos a 42ºC y 10 segundos a 85ºC.

## 3.11 ANÁLISIS FILOGENÉTICOS

*Análisis filogenómicos organulares*

Para realizar los análisis filogenéticos mitocondriales, se concatenaron las secuencias de las proteínas producto de la expresión presentes en los genomas mitocondriales de 25 especies de algas (*cob*, *cox1*, *nad1*, *nad2*, *nad5*, *nad6*) descargados de la base de datos del NCBI con

el programa BIOEDIT (Hall, 1999) y fueron alineadas con el programa MUSCLE (Edgar, 2004) implementado en GENEIOUS R6 (Kearse *et al.* , 2012). Se obtuvo un alineamiento de 2.740 aminoácidos. Las regiones con poca homología entre secuencias fueron eliminadas utilizando el programa GBLOCKS ((Castresana, 2000)) obteniéndose un alineamiento final de 1.706 aminoácidos de los cuales 435 resultaron ser constantes y 1031 fueron informativos para la construcción de la filogenia. Este alineamiento se utilizó como base para la reconstrucción de filogenias utilizando las hipótesis de máxima verosimilitud utilizando el programa PHYML (Guindon & Gascuel, 2003) y de máxima parsimonia utilizando el programa PAUP (Swofford, 2003)

Para la visualización y manipulación de los árboles filogenéticos obtenidos, se utilizó el módulo de python ete2 (Huerta-Cepas *et al.* , 2010) y los programas de manipulación de gráficos vectoriales Inkscape [http://www.inkscape.org/es/] y de mapa de bits GIMP [http://www.gimp.org/].

## 3.12 ANOTACIÓN GENOMA NUCLEAR

Una vez seleccionado el mejor ensamblaje "de novo" realizado con todas las lecturas generadas con el ensamblador Velvet (Zerbino & Birney, 2008), las secuencias obtenidas fueron alineados al conjunto de 458 proteínas presentes en una amplia gama de taxones con el método desarrollado por Parra *et al.* (2009) CEGMA (Core Eukaryotic Genes Mapping Approach). Estas 458 proteínas son genes fundamentales de la base de datos de grupos ortólogos (KOGs) del NCBI de organismos eucariotas, por lo que son útiles para identificar de forma fiable las estructuras exón-intrón de dicho conjunto de genes en las secuencias genómicas de *Trebouxia* sp. TR9. Las búsquedas obtenidas se utilizaron para entrenar al buscador de genes "ab initio" AUGUSTUS (Stanke & Morgenstern, 2005) en su aplicación web (http://bioinf.uni-greifswald.de/webaugustus/index.gsp). El conjunto de modelos obtenido, fue anotado basándose en la similitud de secuencias identificadas utilizando búsquedas BLAST (Altschul *et al.* , 1997) contra la base de datos del NCBI (refseq_protein), el filtrado y anotación funcional como los términos de "Gene Ontology" (GO), los números enzimáticos (EC) o las rutas metabólicas presentes (KEGG) de estas búsquedas fue realizado con Blast2GO (Conesa & Götz, 2008). La herramienta InterProScan-5.10-50.0 (Zdobnov & Apweiler, 2001) fue utilizada para buscar los dominios proteicos de la base de datos PFAM (Bateman *et al.* , 2004) presentes en los modelos proteicos del genoma de *Trebouxia* sp. TR9, en *Asterochloris* sp., *Coccomyxa subellipsoidea*, *Chlorella variabilis*, *Chlamydomonas reindhartii*, *Volvox carteri*, *Micromonas pusilla* CCMP1545, *Micromonas pusilla* RCC299, *Ostreococcus lucimarinus*, *Ostreococcus* sp. RCC809 y *Ostreococcus tauri*, junto a las de un musgo (*Physcomitrella patens*), un helecho (*Sellaginella moellen-*

*dorffii*), una monocotiledónea (*Oryza sativa*) y una dicotiledónea (*Arabidopsis thaliana*) representantes de las embriofitas. Además se realizó la búsqueda de familias relacionadas con el metabolismo de carbohidratos con la herramienta de anotación CAZymes Analysis Toolkit (http://mothra.ornl.gov/cgi-bin/cat/cat.cgi, Park *et al.* 2010), que utiliza la base de datos Carbohydrate Active enZimes (CAZy) (Cantarel *et al.* , 2009) utilizando dos algoritmos diferentes, la búsqueda de secuencias ortólogas con el algoritmo BLAST y la búsqueda de motivos proteicos de la base de datos PFAM en las secuencias analizadas y su conexión con con la correspondiente familia CAZy.

## 3.13 IDENTIFICACIÓN DEL PROTEOMA EXTRACELULAR

Las microalgas *Trebouxia* sp. TR9 y *T. jamesii* son dos microalgas que coexisten en el talo del liquen *Ramalina farinacea*, en el trabajo de Casano *et al.* (2015) se encontraron diferencias en las composiciones de las paredes celulares de estas algas que eran consistentes con sus distintas capacidades para inmovilizar plomo extracelular (Álvarez *et al.* , 2012). Estos autores observaron que los extractos de polímeros extracelulares (EPS) contenían proteínas, y que éstas mostraban patrones diferentes entre *T. jamesii* y *Trebouxia* sp. TR9. Además, observaron que el plomo modulaba el patrón de péptidos extracelulares de *Trebouxia* sp. TR9. En coordinación con el grupo Plantstres del departamento de Ciencias de la Vida de la Universidad de Alcalá, se ha utilizado la anotación del genoma nuclear realizada en la presente Tesis, con el fin de identificar y estudiar las posibles diferencias entre los polipéptidos presentes en los EPS de *Trebouxia jamesii* y *Trebouxia* sp. TR9.

La extracción de las proteínas presentes en las sustancias poliméricas extracelulares de los ficobiontes *Trebouxia* sp. TR9 y *T. jamesii* fue realizada por el grupo Plantstres en la Universidad de Alcalá. Cultivos líquidos monoclonales de los ficobiontes con tres semanas de edad mantenidos en medio 3N-BBM, fueron diluidos a $2x10^6$ células/ml e incubadas con 0 y 100 µM $Pb(NO_3)_2$ disuelto en el medio de cultivo. Los cultivos de las algas con los distintos tratamientos fueron mantenidas en cámara de crecimiento a 15º C bajo ciclos de 14h/10h de luz/oscuridad (25 µmolPAR$m^{-2}s^{-1}$) durante siete días. Tras esto, las algas se centrifugaron (1000 x g, 10 min) y se lavaron con agitación suave con agua ultrapura fría. Se resuspendieron ca. 500 mg de células en 12 ml de agua ultrapura que contenía 0,5 % (v/v) de cóctel de inhibidores de proteasas (Sigma-Aldrich) y 1 mM de PMSF. Las suspensiones de ficobiontes se incubaron a 40ºC durante 40 min con agitación continua y suave. Transcurrido el tiempo, las muestras se centrifugaron a 10.000 x *g* durante 10 min y los sobrenadantes, portadores de las proteínas extracelulares, se filtraron con un filtro de celulosa GF de 10 µm para eliminar posibles células en suspensión.

Las proteínas se dializaron contra agua destilada (24 h, 3 cambios, 4ºC), se liofilizaron y se resuspendieron en agua ultrapura (50-100 µl). La concentración de proteínas en el extracto se valoró con Coomasie Brilliant Blue G-250 siguiendo el ensayo de Bradford (Bradford, 1976) utilizando seroalbúmina bovina como patrón. Las muestras se conservaron a -80ºC para su posterior estudio.

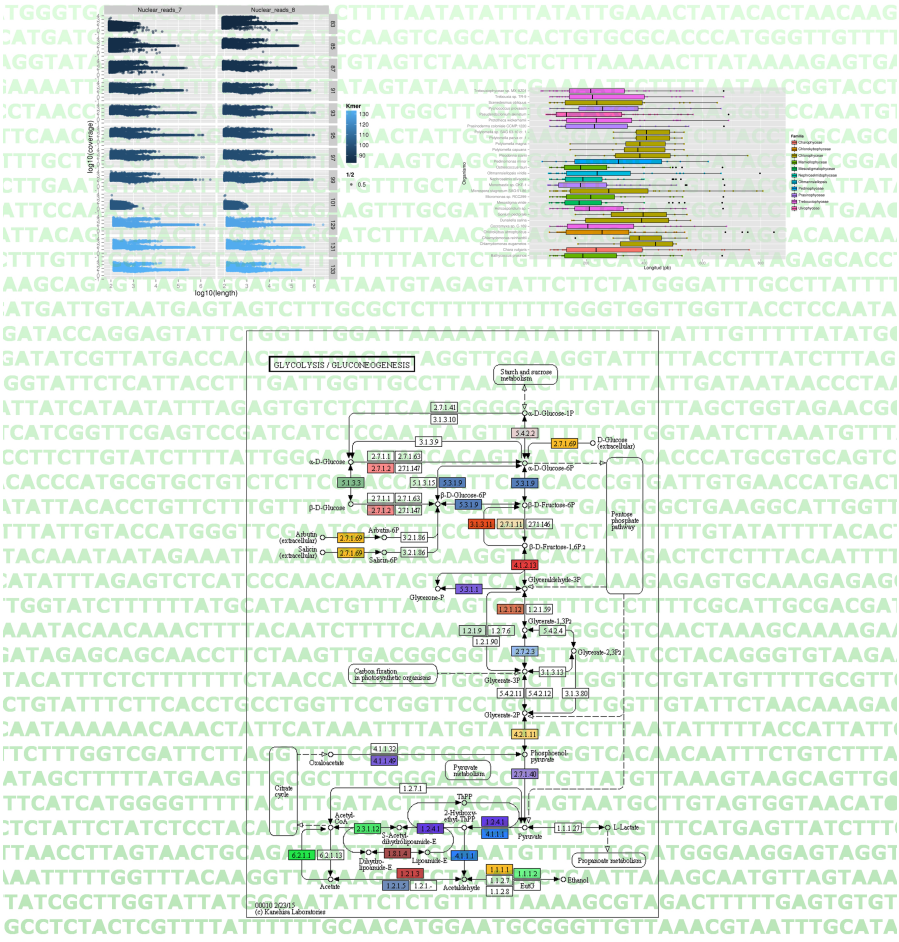
Para el análisis de los exoproteomas de *T. jamesii* y *Trebouxia* sp. TR9 se siguieron dos estrategias experimentales:

1. Secuenciación de los exoproteomas totales obtenidos de los tratamientos control de las dos microalgas llevados a cabo a partir de digestiones en líquido (digeridas sin paso previo de separación en sus diferentes componentes) y analizados mediante cromatografía liquida seguida de una doble espectrometría de masas acoplada a un detector de iones (LC-MSMS). Para la identificación de los espectros de masas obtenidos, se utilizaron dos bases de datos, la base de datos de proteínas del NCBI-Viridiplantae (Noviembre 2014) y los modelos proteicos de *Trebouxia* sp. TR9 obtenidos en el proceso de anotación realizado en esta Tesis.

2. Las exoproteínas de *T. jamesii* y *Trebouxia* sp. TR9, control y tratadas con Pb, también fueron inicialmente separadas mediante electroforesis monodimensionales (SDS-PAGE). Tras la tinción de los polipéptidos con SYPRO® Ruby, se recortaron las bandas más significativas por su abundancia y/o por su variación en respuesta al tratamiento con Pb. Cada una de ellas fue digerida con tripsina y analizada espectrometría de masas acoplada a un detector de iones del tipo "MALDI-TOF". Los espectros obtenidos se identificaron utilizando los modelos proteicos de *Trebouxia* sp. TR9. Los resultados de las identificaciones de los espectros de masas en formato html fueron filtrados con scripts en bash, awk y python para contabilizar y comparar los péptidos identificados en cada alga, tratamiento y base de datos utilizada.

# 4

# RESULTADOS Y DESARROLLO ARGUMENTAL



### 4.1.1 *Resultados*

*Calidad general de la secuenciación y contenido de GC de las lecturas.*

Durante el desarrollo de esta tesis se han realizado tres tipos diferentes de secuenciación masiva de ADN, dos de ellos basados en la tecnología 454 (2010 y 2012) y una en la plataforma Illumina (2014). En la Figura 12 se muestra el resultado del análisis de la calidad de las secuencias de cada tecnología en cada posición nucleotídica. En todos los casos se observa una disminución de la calidad de las secuencias en la parte terminal. Sin embargo, en el caso de las lecturas obtenidas en la plataforma Illumina, la calidad de las lecturas en la parte terminal es mayor que en las obtenidas con la tecnología 454, siendo la secuenciación 454 obtenida en 2010 la que presenta la mayor disminución (Figura 12 A). Consecuentemente, si bien las lecturas obtenidas con la plataforma 454 fueron más largas que las obtenidas en la plataforma Illumina (alrededor de 1200 y 300 pares de bases totales de cada lectura, respectivamente), las lecturas de la plataforma 454 necesitaron ser recortadas para poder trabajar con las zonas de mayor calidad.

Al calcular el contenido en Guanina y Citosina (GC) de dichas lecturas en porcentaje (%GC), se pueden observar dos picos diferentes en el tanto por ciento de Guanina-Citosina (Imágenes internas de la Figura 12). El pico más alto (con mayor abundancia de lecturas) corresponde a las lecturas nucleares y, en todas las secuenciaciones, se encuentra alrededor del 50 %. El pico más bajo (de menor abundancia de lecturas) corresponde a los genomas organulares cuyo %GC se encuentra entorno al 31 %. En el caso de la secuenciación 454 "paired-end" realizada en 2012 (Figura 12 B), estos valores son diferentes por el hecho de contener en el centro de las lecturas la secuencia de unión ("linker") entre ambos extremos "paired-end". Esto se observa en la Figura 13, donde al separar las lecturas en sentido directo y reverso, aparece un pico de alto contenido en GC responsable del aumento en el contenido de GC total correspondiente al adaptador de circularización (Figura 13 C y D).

*Análisis de lecturas obtenidas con la tecnología 454.*

La secuenciación ROCHE 454 GS FLX Titanium del ADN extraído de *Trebouxia* sp. TR9 realizada en 2010, proporcionó un total de 240.256 lecturas (Figura 12 A) con 580 pares de bases de longitud media y una longitud total de 13,92 Mb. Estas secuencias se ensamblaron usando dos ensambladores diferentes (Tabla 6), MIRA V3.2.0 y Newbler 2.6. En el primer caso, se realizaron con cuatro criterios diferentes: tres de ellos sin utilizar un programa externo para filtrar las

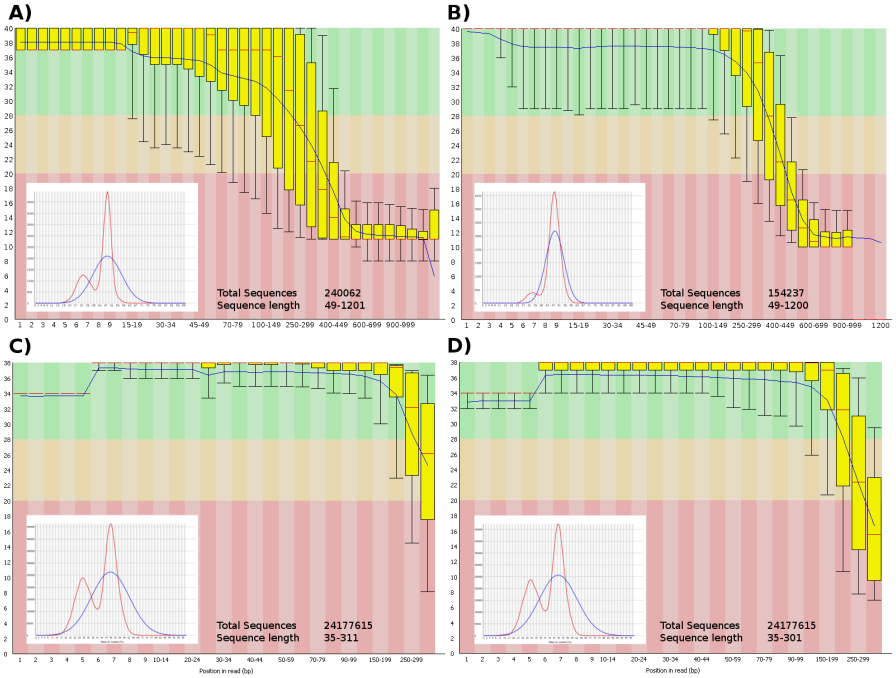



Figura 12: **Calidad de secuencia por nucleótido de lecturas 454 e Illumina secuenciadas (phred score). A)** Lecturas 454 secuenciadas en 2010. **B)** Lecturas 454 "paired-end" (sin separar) secuenciadas en 2012. **C)** Lecturas Illumina en sentido directo secuenciadas en 2014. **D)** Lecturas Illumina en sentido reverso secuenciadas en 2014. Imágenes internas corresponden al contenido en GC de las lecturas. Las líneas roja y azul corresponden al contenido en GC secuenciado y al teórico respectivamente.

secuencias y zonas de baja calidad ("Accurate", "Normal", "Draft") y el cuarto filtrando primero las secuencias con el programa LUCY con la opción "Accurate". Ambos programas filtraron las secuencias de mala calidad tras el ensamblaje, siendo el ensamblaje realizado por MIRA con las lecturas filtradas por LUCY el que más lecturas filtró, pero obtuvo mejores estadísticos tanto para el N50 como para el "contig" más largo. Además, en lo referente al total de nucleótidos ensamblados, se puede observar en la Tabla 6 que tanto Newbler como MIRA + LUCY fueron los ensamblajes más compactos. Esto es debido a que las lecturas obtenidas tenían una baja calidad al final de la secuencia (Figura 12 A) y por tanto en los ensamblajes realizados por MIRA con las lecturas sin filtrar por LUCY, aparecen más "contigs" de menor tamaño debido a problemas de ensamblaje de los errores de secuenciación que posiblemente han llevado a la formación de "contigs" erróneos. En el caso de Newbler, los resultados generales fueron parecidos a los obtenidos por MIRA+LUCY, pero cuando se comparan los "contigs" mayores de 500 pares de bases, frente a un total de nucleótidos ensamblados parecidos (entorno a 450 Kb), el N50 es menor y el número de "contigs" es mayor (Tabla 6).

Cuadro 6: **Estadísticos de Ensamblajes 454**

| | **MIRA V3.2.0** | | | | **Newbler** |
|---|---|---|---|---|---|
| | Accurate | Normal | Draft | Luce Accurate | 2010 |
| Lecturas iniciales | 240.256 | 240.256 | 240.256 | 233.065 | 240.097 |
| Lecturas ensambladas | 182.803 | 182.309 | 180.924 | 140.740 | 155.805 |
| **"contigs" totales** | | | | | |
| Nº "contigs" | 20.721 | 20.787 | 21.014 | 6.757 | 9.300 |
| Total (nt) | 9.190.616 | 9.211.887 | 9.254.291 | 4.711.981 | 5.916.780 |
| Largest "contig" | 7.713 | 7.499 | 6.754 | 59.973 | 21.725 |
| N50 | 517 | 516 | 509 | 729 | 677 |
| **Large "contigs"** | | | | | |
| Nº "contigs" | 228 | 285 | 366 | 123 | 6.148 |
| Total (nt) | 459.291 | 471.501 | 501.042 | 435.925 | 475.028 |
| N50 | 2.619 | 1.876 | 1.442 | 7.966 | 751 |

La secuenciación ROCHE 454 "paired-end" GS JUNIOR del ADN extraído de *Trebouxia* sp. TR9 realizada en el año 2012, proporcionó un total de 154.237 lecturas (Figura 12) con 393 pares de bases de longitud media y una longitud total de 60,66 Mb. Estas lecturas contenían sin separar las lecturas en sentido directo y reverso unidas entre sí por una secuencia "linker" con la que se circularizó la molécula antes de añadirle los adaptadores de secuenciación (Ver Sección 1.5). Cuando estas lecturas fueron separadas entre sí y del adaptador, se obtuvieron 104.113 secuencias apareadas tanto en sentido directo como reverso con 191 y 175 pares de bases de longitud media junto a un total de 19,95 y 18,22 nt, respectivamente. Además, se obtuvieron 49.117 secuencias sin su pareja complementaria con 349 nt de longitud media y una longitud total de 17,17 Mb. Estas lecturas aparecieron por dos posibles causas: debido a la baja calidad de la parte 3' o debido a errores de secuenciación en la parte del adaptador por lo que el software utilizado para separarlas, no reconoció dicha secuencia y por tanto no separó las lecturas entre sí del "linker".

Las lecturas 454 "paired-end" obtenidas en 2012 fueron ensambladas junto a las obtenidas en 2010 con el ensamblador Newbler. Este

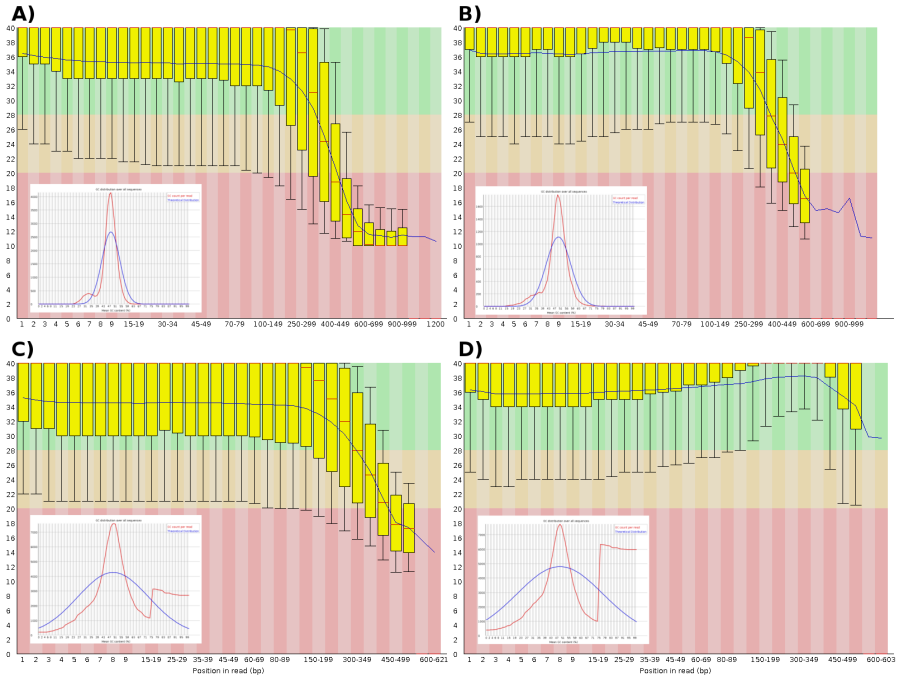



Figura 13: **Calidad de secuencia por nucleótido de lecturas 454 "paired-end" secuenciadas en 2012. A)** Lecturas 454 "paired-end" sin separar. **B)** Lecturas 454 "paired-end" en que al ser separadas, no obtuvieron la secuencia asociada en el otro sentido. **C)** Lecturas 454 paired-end en sentido directo que al ser separadas tenían asociado el extremo reverso. **D)** Lecturas 454 paired-end en sentido reverso que al ser separadas tenían asociado el extremo directo. Imágenes internas corresponden al % de GC de las lecturas. Las líneas roja y azul corresponden al % de GC secuenciado y al teórico, respectivamente.

ensamblador fue seleccionado debido a que utiliza la información de los extremos apareados secuenciados para formar "scaffolds". Se probaron diferentes valores para el parámetro *overlapMinMatchIdentity* y el mejor ensamblaje fue seleccionado comparando los valores de la Tabla 7. Para realizar esta selección se utilizó como referencia la secuencia del genoma mitocondrial secuenciado anteriormente (Ver sección 4.2 de Resultados). En todas las opciones probadas se obtuvo que el "contig" más largo de los "scaffolds" formados pertenecía a este genoma, menos en la de valor 100. Por tanto, se descartó el ensamblaje realizado con la opción de similitud entre lecturas de valor 100. Al comparar los ensamblajes de las demás opciones entre sí, todos los valores eran muy semejantes a excepción del tamaño del "scaffold" más grande, donde la opción 90 obtuvo el tamaño mayor y por lo tanto, fue seleccionado para, posteriormente, obtener el genoma cloroplástico (Ver sección 4.3) o para guiar al ensamblador Velvet a realizar el ensamblaje de las lecturas Illumina obtenidas en 2014 (Ver más abajo).

Cuadro 7: **Estadísticos de los ensamblajes realizados con la opción del ensamblador Newbler *overlapMinMatchIdentity* (Tamaños en Kilo bases)**

| **Assembly option** | **80** | **85** | **90** | **95** | **100** |
|---|---|---|---|---|---|
| Tamaño del "contig" Mayor de los "Scaffolds" | 70,07 | 70,074 | 70,077 | 70,076 | 7,922 |
| Tamaño "Scaffold" Mayor | 138,147 | 140,42 | 272,617 | 125,610 | 39,901 |
| N50 de "contigs" | 3,385 | 3,645 | 3,458 | 3,678 | 3,013 |
| N50 de "Scaffolds" | 4,077 | 4,080 | 4,099 | 4,122 | 9,245 |
| Nucleótidos de "Scaffolds" | 1,025 | 1,014 | 1,016 | 1,008 | 0,395 |
| Nucleótidos Totales | 10,326 | 10,325 | 10,326 | 10,261 | 6,619 |

*Análisis de las lecturas obtenidas con la plataforma Illumina.*

La secuenciación Illumina Miseq paired-end del ADN extraído de *Trebouxia* sp. TR9 en el año 2014, proporcionó más de 24 millones de lecturas tanto en sentido directo como reverso (Figura 12 C y D). En esta secuenciación se generó una librería con un rango de tamaños de inserto comprendido entre 400 y 2000 nt. El 94 % de los fragmentos de la librería generada se encontraba entre tamaños de 610 y 1400 pares de bases (Figura 14). Las lecturas obtenidas, tanto en sentido directo como reverso, tenían una longitud de 311 y 301 pares de bases y un total de 7.507 y 7.265 Mb respectivamente.

Las lecturas fueron filtradas y recortadas las zonas de mala calidad con la opción [*Number of passes=0*] del ensamblador MIRA4. El número de lecturas que, superaron los procesos de recorte por calidad y aquellas parejas de lecturas que pasaron dicho filtrado, fue de 16.461.711 en cada orientación. Este grupo de lecturas fueron ensambladas "de novo" con el ensamblador Velvet con diferentes tamaños de palabra ("K-mer"). Un total de 57 ensamblajes fueron realizados con tamaños de "K-mer" impares desde el 21 al 133. Para estimar el mejor ensamblaje, en primera estancia se compararon los logaritmos en base 10 de las coberturas (en "K-mers") frente al logaritmo en base 10 de las longitudes de cada "contig" para cada ensamblaje. En la Figura 15 se puede observar que los ensamblajes realizados con

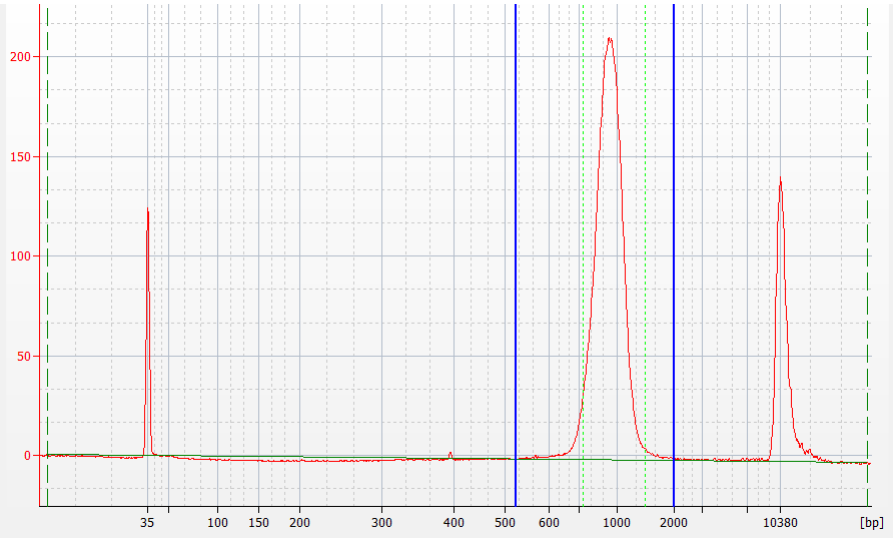



Figura 14: **Distribución de tamaños de la librería Illumina paired-end secuenciada.**

tamaños de “K-mer” del 21 - 79 y del 101 - 127 dieron ensamblajes donde la continuidad de los "contigs" era muy baja; además, cuando fueron contados el número de nucleótidos marcados con una “N”, los ensamblajes con un tamaño de palabra comprendidos entre 21 - 33 y 101 - 125 no produjeron ningún nucleótido marcado con una “N”, por lo que no se formó ningún “scaffold” y estos ensamblajes fueron descartados.

Una vez filtrados estos ensamblajes, se procedió a comprobar el porcentaje de genes nucleares conservados en eucariotas (CEGS) presentes en los ensamblajes restantes con tamaño de palabra entre 83 - 101 y 127 - 133. El espacio génico, es una medida útil para describir lo completo que es un ensamblaje. En la Figura 16 se representa la longitud N50 frente al porcentaje de CEGS encontrados en cada ensamblaje. En este caso se añadieron los ensamblajes realizados con un “K-mer” de 101 y 127, antes filtrados, como controles negativos para comprobar la metodología.

Los ensamblajes con un “K-mer” de 95 y 97 obtuvieron el número más alto de CEGs completos y parciales (Tabla 8). Finalmente se seleccionó el ensamblaje realizado con el “K-mer” 97 puesto que ante valores similares de N50, N95 y el tamaño total de residuos ensamblados, poseía un menor número de "contigs"/“scaffolds”, por lo que este ensamblaje fue utilizado para analizar el genoma nuclear al encontrarse menos fragmentado.

### 4.1.2 *Discusión*

El ensamblaje de genomas a partir de las lecturas obtenidas con las tecnologías de secuenciación masiva, sigue siendo uno de los proble-

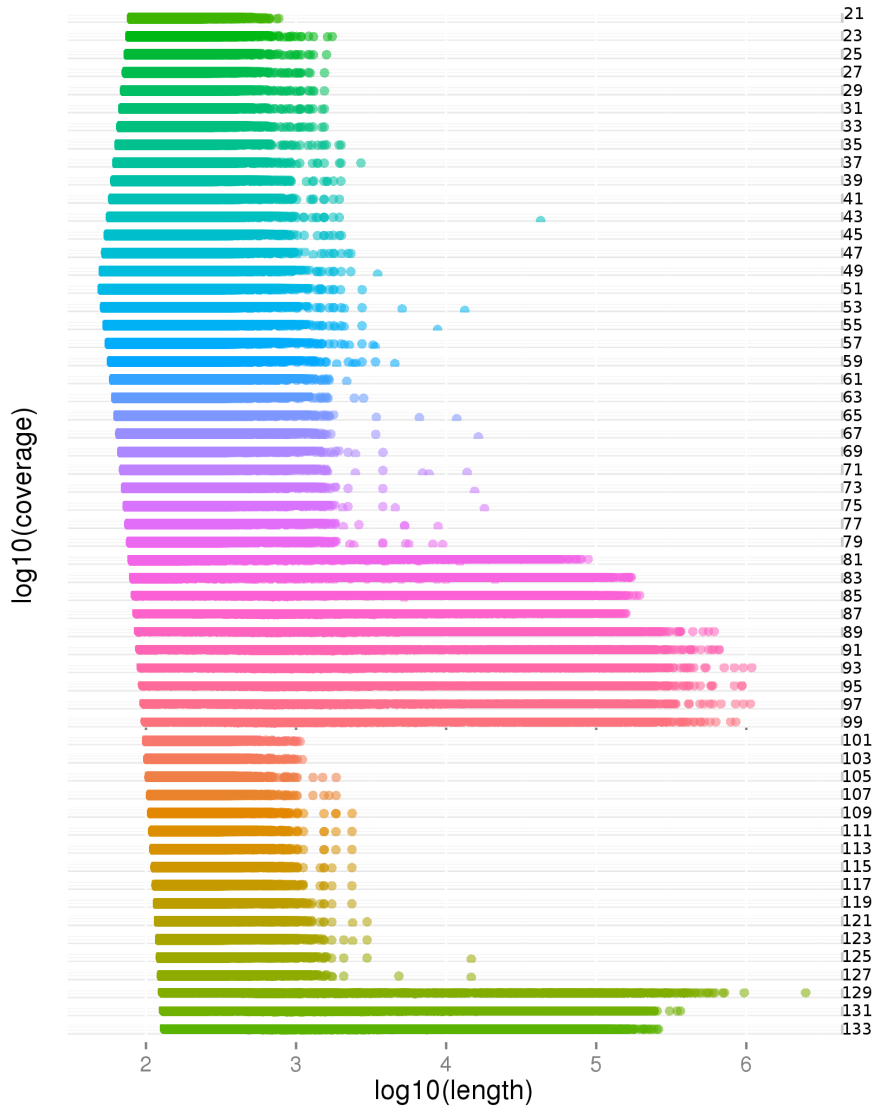



Figura 15: **Valores de Log10 (cobertura) frente al Log10 (longitud) de los "contigs" de los ensamblajes realizados con Velvet.**

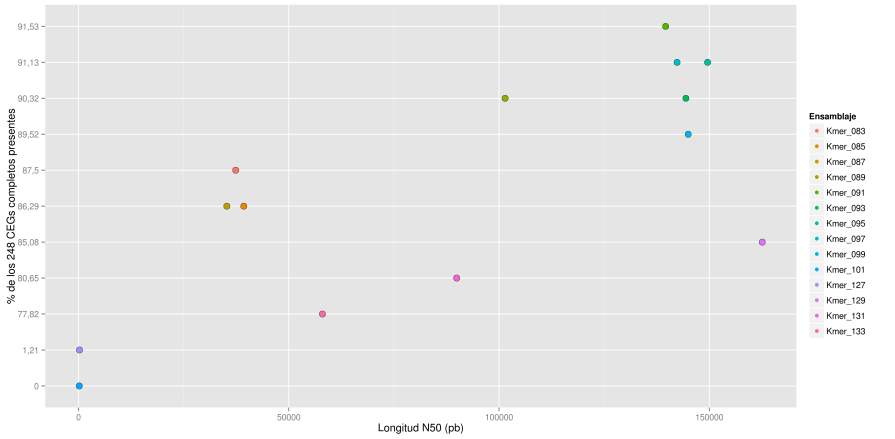



Figura 16: **Valores de porcentaje de genes nucleares conservados en organismos eucariontes (CEGS) frente a valores de la longitud N50.**

mas centrales de la bioinformática. Esto se debe, en gran medida, a las diferencias en las propiedades de las lecturas obtenidas como son la longitud y los errores de secuenciación. Por otra parte, la existencia de regiones repetidas en los genomas plantean nuevos retos para el correcto ensamblaje. Existen una serie de ventajas e inconvenientes aparentes dependiendo del tipo de tecnología de secuenciación. Generalmente, cuanto mayor es la longitud de lecturas obtenidas, mejor se resuelven las repeticiones presentes en el genoma a secuenciar, pero al mismo tiempo, ello implica un mayor número de errores al final de las lecturas.

En el transcurso de la presente tesis se han obtenido lecturas de dos plataformas diferentes de secuenciación, ROCHE 454 e Illumina. En el primer caso, se realizaron dos librerías diferentes, una en el año 2010 bajo la plataforma ROCHE 454 GS FLX Titanium con lecturas en sentido directo y otra en el año 2012 bajo la plataforma ROCHE 454 "paired-end" GS JUNIOR, donde se utilizó una librería "paired-end" con un tamaño de inserto entre ambos extremos de 3 kb. En el caso de la secuenciación Illumina MiSeq "paired-end", se utilizó una librería con tamaños de inserto comprendidos entre 610 y 1400 pares de bases.

Al comparar las lecturas obtenidas por cada una de las plataformas, observamos que las pertenecientes a la tecnología 454 proporcionaban lecturas de tamaños mayores, pero con calidades menores al final de las secuencias. Por este motivo fue necesario filtrarlas por calidad y recortarlas en tamaños menores (alrededor de 450 pares de bases). Además, en el caso de la librería "paired-end", al separar las lecturas en sentido directo y reverso quitando la secuencia del adaptador de circularización, las secuencias en sentido reverso resultaron más cortas debido a la existencia de zonas de baja calidad al final de las lecturas, e incluso algunas no pasaron el filtrado por calidad/longitud, quedando tan solo la secuencia en sentido directo. En cambio,

Cuadro 8: **Estadísticos de los mejores ensamblajes realizados con Velvet.**

| K-mer | N50 | N95 | Residuos (Mb) | Nº "contigs" | Nº N's | % CEGs Completos | % CEGs Parciales |
|---|---|---|---|---|---|---|---|
| **83** | 37.383 | 1.239 | 59,82 | 13.531 | 2.612 | 87,50 | 96,37 |
| **85** | 39.303 | 2.424 | 59,13 | 9.252 | 6.686 | 86,29 | 95,97 |
| **87** | 35.277 | 2.714 | 59,16 | 8.265 | 50.721 | 86,29 | 95,56 |
| **89** | 101.444 | 6.530 | 59,42 | 4.913 | 381.067 | 90,32 | 9,58 |
| **91** | 139.611 | 13.422 | 59,45 | 3.703 | 420.747 | 91,53 | 97,58 |
| **93** | 144.417 | 16.777 | 59,55 | 3.260 | 501.263 | 90,32 | 97,18 |
| **95** | 149.579 | 18.804 | 59,53 | 2.995 | 531.286 | 91,13 | 97,18 |
| **97** | 142.331 | 19.150 | 59,50 | 2.899 | 490.261 | 91,13 | 97,18 |
| **99** | 144.973 | 19.529 | 59,47 | 2.797 | 457.885 | 89,52 | 96,77 |
| **101** | 201 | 201 | 321,84 | 1.573.598 | 0 | 0,00 | 1,61 |
| **127** | 253 | 253 | 156,67 | 582.278 | 183 | 1,21 | 6,45 |
| **129** | 162.583 | 22.177 | 61,69 | 2.412 | 797.223 | 85,08 | 94,76 |
| **131** | 89.919 | 14.100 | 74,66 | 4.136 | 2.428.712 | 80,65 | 94,35 |
| **133** | 58.008 | 5.233 | 79,08 | 7.187 | 2.479.805 | 77,82 | 92,34 |

las lecturas pertenecientes a la tecnología Illumina presentaron una calidad mayor, por lo que en el proceso de filtrado se mantuvieron mucho más las longitudes iniciales. En lo relativo al número de secuencias, las lecturas pertenecientes a la plataforma Illumina dieron un rendimiento muchísimo mayor, permitiendo ensamblar una gran parte (si no el total) del genoma nuclear.

En el proceso de ensamblaje de las lecturas obtenidas, se han utilizado tres ensambladores diferentes: MIRA, GsAssembler (Newbler) y Velvet. Cada ensamblador presentó diferentes ventajas e inconvenientes. El ensamblador MIRA resultó muy eficaz debido a que utiliza un algoritmo basado en alineamientos Smith-Waterman, y por tanto, mantiene la información total de las lecturas, resolviendo así muchas repeticiones presentes en el genoma. Además, fue muy útil en el pre-procesado y filtrado de calidad de lecturas de tipo Illumina. Sin embargo, cuando se trató de procesar una gran cantidad de datos como los proporcionados por la plataforma Illumina, el enorme uso que hace este programa de la memoria y el espacio de disco duro, imposibilitó su utilización. Por otra parte, es incapaz de formar "scaffolds" a partir de datos "paired-end". El ensamblador Newbler que es específico para la plataforma 454, es capaz de filtrar por sí solo las lecturas y formar "scaffolds" utilizando la información de las lecturas "paired-end". Sin embargo, uno de los grandes inconvenientes que

presenta, aparte de ser de código cerrado y por tanto los algoritmos que utiliza son una "caja negra", es que en el caso de repeticiones, los algoritmos internos de este ensamblador provocan la pérdida de la información de continuidad de las lecturas y por tanto no resuelve adecuadamente estas repeticiones. El ensamblador Velvet, que utiliza grafos de Bruijn, aceptó todas las lecturas obtenidas, tanto de 454 como de Illumina, con lo que se obtuvieron buenos rendimientos en términos de velocidad y longitud de "contigs" / "scaffolds". Uno de los inconvenientes de este ensamblador a la hora de resolver el grafo de Bruijn, es el uso de paralelizaciones de CPU. Frente al incremento de velocidad para los ensamblajes, Velvet produce diferentes resultados a partir de los mismos datos y opciones, por lo que impide la reproducibilidad de los ensamblajes. Una solución a este hecho es utilizar tan solo una CPU, con lo que el proceso de ensamblaje tardaría mucho más, pero los resultados serían reproducibles.

El uso de las tecnologías de secuenciación masiva ha abierto una nueva puerta a la secuenciación de genomas. Los desarrollos tecnológicos en la generación más rápida y simplificada de librerías, el aumento tanto de la longitud de las lecturas como de la calidad de las mismas o la aparición de algoritmos más potentes y eficientes en el uso de recursos informáticos, esta ayudando a responder importantes preguntas sobre evolución, filogenia y taxonomía. Aún así, es necesario el uso de diferentes herramientas bioinformáticas para poder filtrar los errores de secuenciación y mejorar el manejo de las repeticiones presentes en los genomas. Para ello, es aconsejable utilizar dos o más tipos de secuenciaciones (En este caso 454 e Illumina) y así corregir los errores de cada tecnología, junto con la utilización de tamaños diferentes de librerías "paired-end" para mejorar la conectividad de "contigs" y la formación de "scaffolds".

## 4.2 GENOMA MITOCONDRIAL

### 4.2.1 *Resultados*

Se ha obtenido la secuencia completa del genoma mitocondrial de *Trebouxia* sp. TR9. Las lecturas generadas en la primera secuenciación realizada con la tecnología de pirosecuenciación 454 (ROCHE 454 GS FLX Titanium) se ensamblaron con el programa MIRA a partir de las lecturas filtradas con LUCY, ya que, como se explicó en el apartado 4.1.1 de Resultados, fue la combinación que propició el mejor ensamblaje. Con el objeto de obtener el genoma mitocondrial completo, se han seleccionado los "contigs" mitocondriales con diferentes búsquedas con el algoritmo BLAST. De este modo, se filtraron dos grandes "contigs" de 10 y 60 kb que fueron unidos gracias al diseño de oligos específicos, amplificación por PCR y secuenciación con el método de Sanger. De esta manera se obtuvo una secuencia única circular de

más de 70 kb correspondiente al genoma mitocondrial de *Trebouxia* sp. TR9. Posteriormente, se obtuvieron nuevas lecturas 454 "paired-end" (ROCHE 454 "paired-end" GS JUNIOR) tal como se describió en el apartado 4.1.1 de Resultados. Estas nuevas lecturas se mapearon utilizando como secuencia molde el genoma mitocondrial obtenido, confirmando así el ensamblaje obtenido (Figura 17). El correcto posicionamiento de las lecturas mapeadas ha servido para poder estimar que el tamaño medio de inserto de la librería era de 1.800 nt con una desviación de 563,1 nt y también para confirmar que el genoma mitocondrial de *Trebouxia* sp. TR9 es circular.

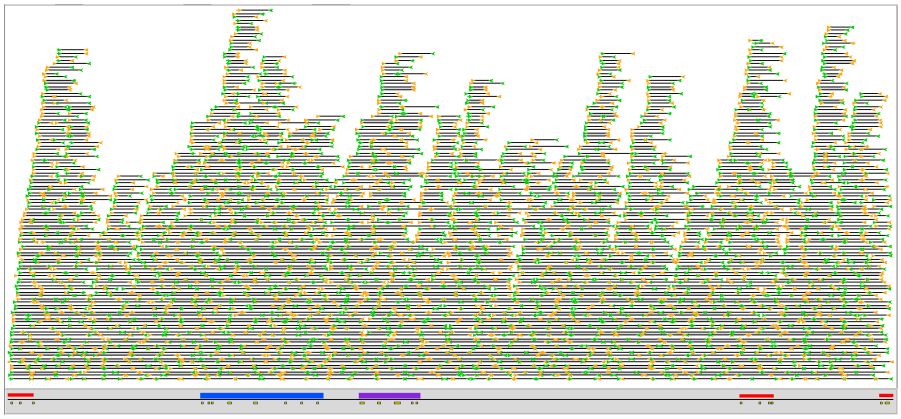


Figura 17: **Mapeo de lecturas 454 "paired-end" contra el genoma mitocondrial de *Trebouxia* sp. TR9.** Las flechas de colores unidas por una línea negra en la parte superior son las parejas de lecturas mapeadas. En la parte inferior se muestra un esquema del genoma mitocondrial (línea negra), la posición de las PCR's de unión (rectángulos rojos), el gen *cox1* (rectángulo morado), el gen *rrnL* (rectángulo azul) y la posición de los oligos diseñados para realizar PCR's y RNA amplificado con retro-transcriptasa (cuadrados amarillos).

*Tamaño del genoma, estructura y genes codificados.*

Los resultados obtenidos indican que la molécula del ADN mitocondrial de *Trebouxia* sp. TR9 es de 70.070 nt (Figura 18). La longitud total de los genes mitocondriales de *Trebouxia* sp. TR9 asciende a 31.118 nt, quedando como zonas no codificantes 38.951 nucleótidos del total (70.070 nt). Este dato indica que aproximadamente el 50 % del genoma mitocondrial de *Trebouxia* sp. TR9 está ocupado por secuencias no codificantes. Los genes codificados en el genoma mitocondrial de *Trebouxia* sp. TR9 presentan una longitud de 555 nt y un 39 % de contenido medio de GC. En las regiones intergénicas del genoma mitocondrial de *Trebouxia* sp. TR9 aparecen 15 elementos repetitivos que suman aproximadamente el 1,88 % del total de nucleótidos, su contenido en GC es del 12 al 57 % (40 % de media) y miden

de 28 a 175 nt de longitud (45 nt de media). De ellos, 9 son secuencias invertidas y 6 se disponen en sentido directo (Figura 18).

En la secuencia obtenida, se han identificado un total de 61 genes (Tabla9). Dieciocho de estos genes codifican proteínas implicadas en el metabolismo energético de la mitocondria, 13 codifican proteínas ribosomales (4 de la sub-unidad grande y 9 de la pequeña), uno de ellos codifica una proteína implicada en el transporte transmembrana de proteínas, 29 codifican ARNs estructurales (3 codifican ARNs ribosomales y 26 codifican ARNs de transferencia).

*Presencia de intrones en genes codificantes de proteínas de Trebouxia sp. TR9.*

En algunos de los genes identificados se han localizado un total de nueve intrones (Tabla 9). Todos ellos son del tipo I, siendo más abundantes los pertenecientes al grupo IB presentes en genes que codifican tanto proteínas (cox1) como ARNs (*rrnL*), seguidos de los del grupo IA restringidos a genes que codifican ARNs (*rrnL* y *rrnS*) y un único intrón perteneciente al grupo ID en el gen *cob*. El tamaño de los intrones es variable, desde 500 nt (el único intrón del gen *rrnS*) hasta 1.443 nt (el tercer intrón del gen *cox1*). Solamente los intrones presentes en los genes que codifican proteínas, *cox1* y *cob*, contienen ORFs que codifican "Homing endonucleases" de la familia LAGLIDADG (LHEs), con uno o dos motivos en cada una de ellas.

Al comparar con la herramienta BLAST las secuencias de los intrones de *Trebouxia* sp. TR9 con las depositadas en GenBank, encontramos dos resultados inesperados. Por una parte, el intrón del gen *cob* presenta unos valores elevados de identidad de secuencia al compararlo con el correspondiente de *Chlorokybus atmophyticus* (acceso EF463011), la cual es un alga verde de la división Streptophyta que junto a *Mesostigma viride* están en la base de la divergencia de esta división frente a la de Chlorophyta *(Turmel* et al. , *2007)*. El porcentaje de secuencia alineada del intrón que presentan ambos organismos es del 65 % y el % de identidad de nucleótidos es del 81 % (e-value=$3e^{-58}$). Si hacemos la misma comparación con *Trebouxia aggregata* (acceso EU123948) se obtienen valores del 76 % y 77 %, respectivamente (e-value=$9e^{-33}$). Por otra parte, al comparar la secuencia de nucleótidos del primer intrón del gen *cox1* con secuencias de otros organismos del GenBank, encontramos las similitudes más significativas con intrones del mismo gen de hongos. Por ejemplo, valores de 83 % de cobertura de la secuencia del intrón y 65 % de identidad de nucleótidos (e-value=$2e^{-41}$) cuando se compara con el basidiomiceto *Trametes cingulata* (acceso GU723273). Resultados similares se obtuvieron con otros hongos. La Figura 19 muestra los alineamientos de las secuencias de las LHEs codificadas en los intrones de los genes *cob* y *cox1*. Estos resultados señalan la complejidad del origen y evolución de los intrones de tipo I y las LHEs codificadas en ellos, que frecuentemente

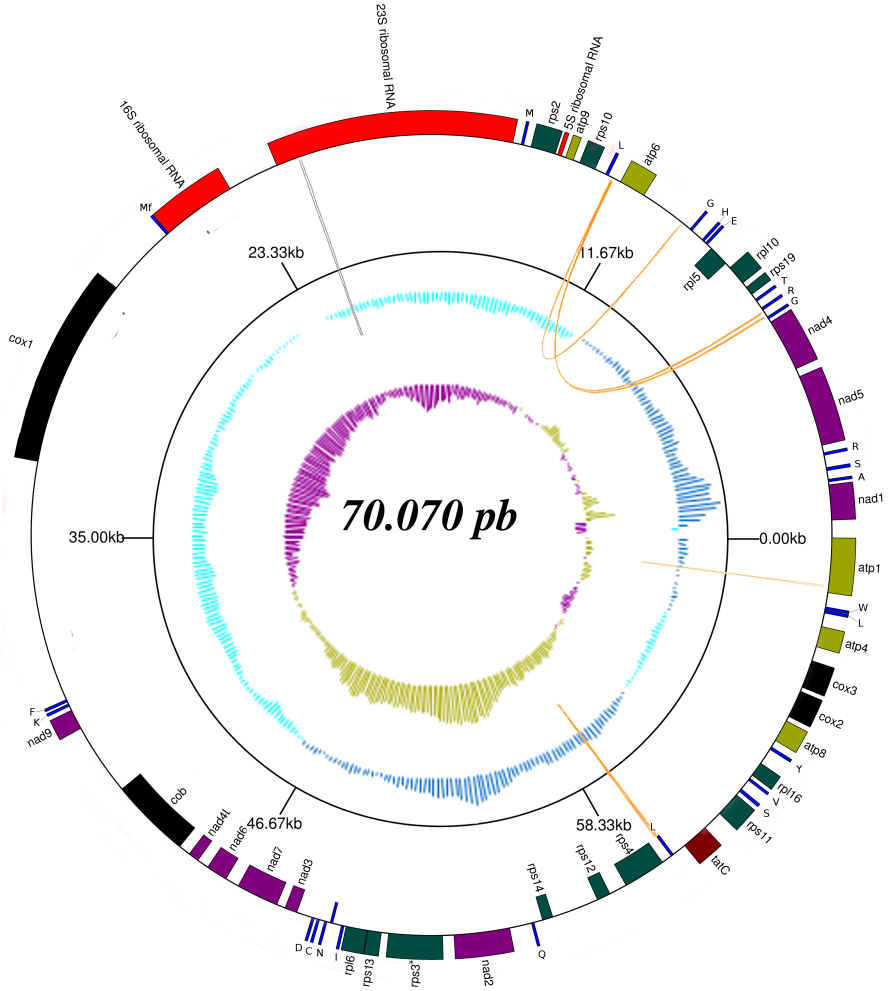



Figura 18: **Mapa del genoma mitocondrial de *Trebouxia* sp. TR9.** Los diferentes anillos, desde el interior hacia el exterior, muestran el sesgo en GC (GC skew), el contenido en GC, los tamaños en Kb y los genes como cajas de colores. Cuando las cajas se sitúan fuera y dentro del anillo, los genes se expresan en sentido horario y antihorario respectivamente. En rojo se presentan los genes relacionados con la fosforilación oxidativa, en naranja las proteínas ribosomales, en azul claro los ARNs de transferencia y en azul oscuro los genes ribosomales. Los genes situados fuera y dentro del círculo se expresan en sentido horario y antihorario respectivamente. Lineas grises y naranjas en el fondo unen las repeticiones presentes en el genoma.

Cuadro 9: **Genes identificados en el genoma mitochondrial de *Trebouxia* sp. TR9.**

| Función | Número de genes | Número de intrones | Nombre del gen |
|---|---|---|---|
| **Complejo III / IV** | | | |
| Citocromo b/ Oxidasas | 4 | 4 | *cob*[a], *cox1*[a], *cox2*, *cox3* |
| **Complejo V** | | | |
| ATP sintasa | 5 | 0 | *atp1*, *atp4*, *atp6*, *atp8*, *atp9* |
| **Complejo I** | | | |
| NADH deshidrogenasa | 9 | 0 | *nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7*, *nad9* |
| **Proteínas ribosomales** | | | |
| Subunidad grande | 4 | 1 | *rpl5*, *rpl6*, *rpl10*, *rpl16* |
| Subunidad pequeña | 9 | 0 | *rps2*, *rps4*, *rps7*, *rps10*, *rps11*, *rps12*, *rps13*, *rps14*, *rps19* |
| **Proteína transportadora de membrana** | 1 | 0 | *mttB* (*TatC* en *E. coli*) |
| **RNAs ribosomales** | 3 | 5 | *rrnL*[a], *rrnS*[a], *rrnF* |
| **RNAs de transferencia** | 26 | 0 | *trnA* (UGC), *trnS* (GCU), *trnR* (UCU), *trnG* (GCC), *trnR* (ACG), *trnT* (UGU), *trnE* (UUC), *trnH* (GUG), *trnG* (UCC), *trnL* (UAG), *trnMe* (CAU), *trnMf* (CAU), *trnF* (GAA), *trnK* (UUU), *trnD* (GUC), *trnC* (GCA), *trnN* (GUU), *trnI* (CAU), *trnI* (GAU), *trnQ* (UUG), *trnL* (CAA), *trnS* (UGA), *trnV* (UAC), *trnY* (GUA), *trnL* (UAA), *trnW* (CCA) |

[a] Genes que contienen intrones

están sujetos a procesos de transferencia horizontal entre diferentes genomas y organismos (Friedl *et al.* , 2000; del Campo *et al.* , 2009).

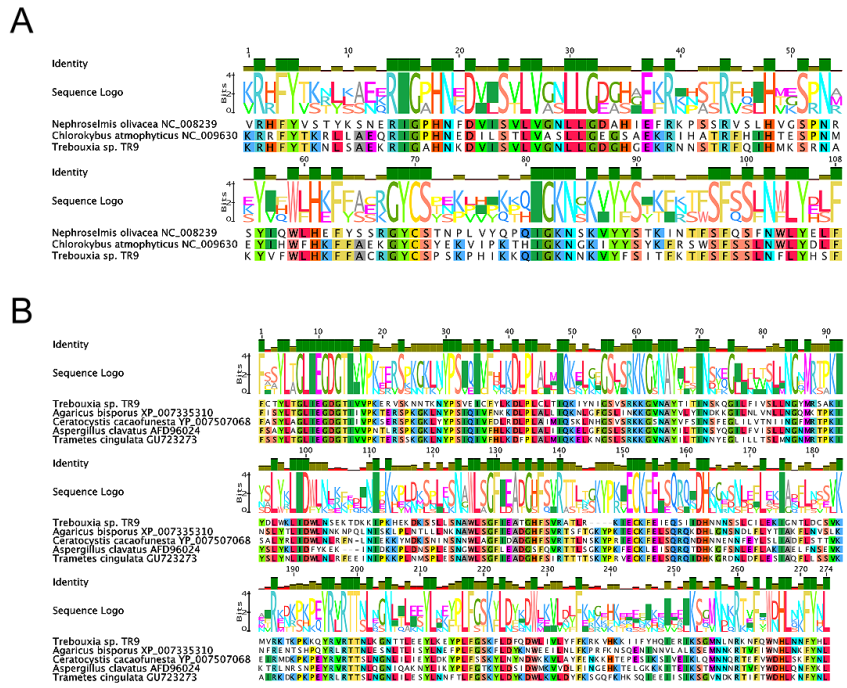



Figura 19: **Alineamiento de secuencias de aminoácidos de Homing endonucleasas de tipo LAGLIDADG (LHEs) de *Trebouxia* sp. TR9 con LHEs similares de otros organismos.** Codificadas en los intrones de los genes *cob* (A) y *cox1* (B)

En el genoma mitocondrial de *Trebouxia* sp. TR9 no se han identificado intrones de tipo II enteros, que sí están presentes en los genomas mitocondriales de otras algas de la división Chlorophyta, incluso pertenecientes a la clase Trebouxiophyceae, como por ejemplo *Coccomyxa subellipsoidea* C-169 (acceso NC_015316). Sin embargo, en el genoma mitocondrial de *Trebouxia* sp. TR9 se ha identificado una pequeña secuencia de 75 nt entre los genes *trnG* (UCC) y *atp6* (224 nt por delante de este último) que se corresponde con los dominios V y VI propios de intrones de tipo II. Sin embargo, en la Figura 20 puede observarse la extraordinaria conservación de la estructura secundaria de los dos dominios V y VI así como de motivos importantes para el mecanismo de "autosplicing" en intrones de tipo II. También puede observarse una gran similitud, tanto en secuencia como en estructura secundaria, con los dominios V y VI del intrón de tipo II B1 presente en el gen *rrnS* en la posición 788 (SSU788) del basidiomiceto *Grifola frondosa* (Figura 20).

*Preferencia de codones y código genético.*

El programa FACIL (Dutilh *et al.* , 2011) permite la predicción automática del código genético a partir de una secuencia o grupo de

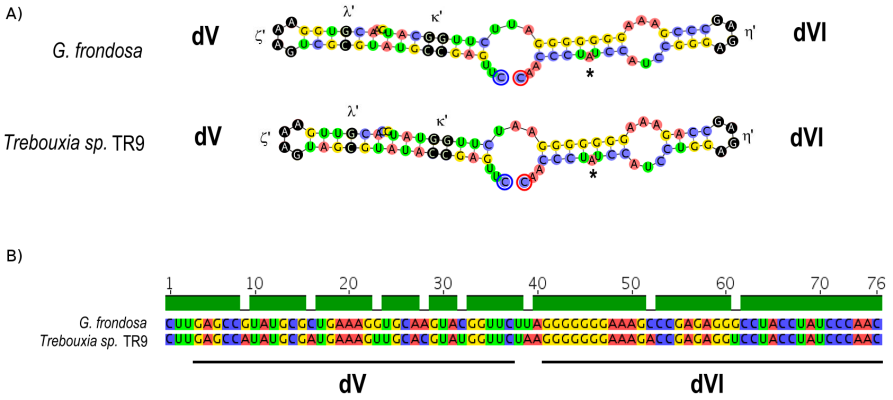



Figura 20: **Secuencia y estructura secundaria de los dominios V y VI propios de intrones de tipo II de *Grifola frondosa* y *Trebouxia* sp. TR9.** (A) Estructuras secundarias y motivos importantes para el "autosplicing" según Michel *et al.* (2009). La "A" protuberante ("bulging A") se señala con un asterisco. (B) Alineamiento de las secuencias de *Grifola frondosa* y *Trebouxia* sp. TR9.

secuencias de nucleótidos del genoma cuyo código se quiere determinar. El programa busca qué aminoácidos en proteínas homólogas de otros organismos se alinean con mayor frecuencia en cada codón. La probabilidad resultante de aminoácidos para cada codón es mostrada como un "logo" del código genético que también resalta los codones de terminación de la traducción. El programa FACIL predijo el "código estándar" (NCBI translation table 1) como el código más similar en el genoma mitocondrial de *Trebouxia* sp. TR9, con 62 de 64 codones correctamente predichos (96,9 %). Curiosamente, la predicción para los códigos 12 y 16 fue de un 95,3 %. La predicción para el código 22 fue de 93,8 %. Las predicciones para los demás códigos obtuvieron valores inferiores. En la Figura 21 se muestra el "logo" de la frecuencia de cada aminoácido correspondiente a cada uno de los codones posibles, incluidos los de terminación de la traducción para *Trebouxia* sp. TR9, *Arabidopsis thaliana* (acceso NC_001284), *Chlamydomonas reinhardtii* (acceso NC_001638) y *Scenedesmus obliquus* (acceso NC_002254), cuyos genomas mitocondriales utilizan los códigos "estándar", "de mitocondrias de algas de la clase Chlorophyceae" y "de Scenedesmus" respectivamente. Como puede observarse en la Figura 21, el código genético utilizado en la mitocondria de *Trebouxia* sp. TR9 es muy similar al de *A. thaliana* que utiliza el código genético "estándar" y menos parecido al de *C. reinhardtii* y *S. obliquus*. Este resultado nos ha llevado a adoptar el código genético "estándar" como el adecuado para la expresión del genoma mitocondrial de *Trebouxia* sp. TR9.

Complementariamente, la Figura 22 muestra la frecuencia de los aminoácidos presentes en el conjunto de los genes que codifican proteínas en el genoma mitocondrial de *Trebouxia* sp. TR9. Los aminoáci-

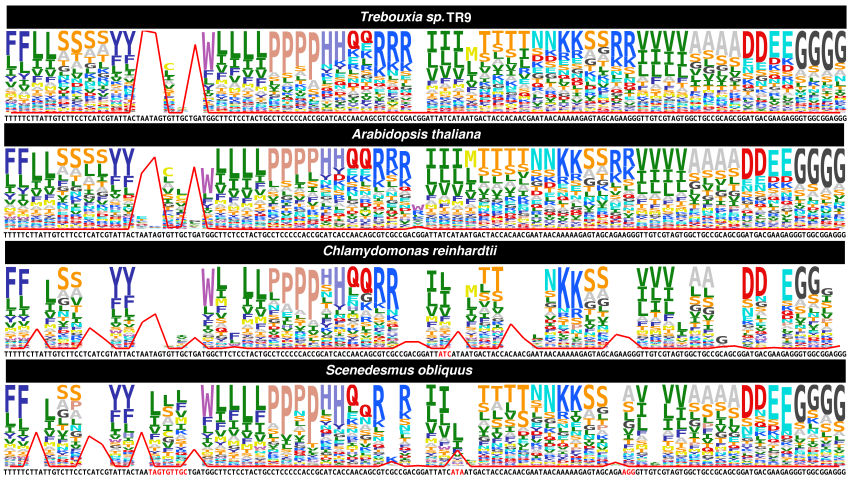



Figura 21: **Esquema del "logo" de la frecuencia de cada aminoácido correspondiente a cada uno de los codones posibles para los genomas mitocondriales de la planta vascular *Arabidopsis thaliana* y las algas *Trebouxia* sp. TR9, *Chlamydomonas reinhardtii* y *Scenedesmus obliquus*.** La línea roja representa los codones de terminación de la traducción.

dos hidrofóbicos leucina (L), isoleucina (I) y fenilalanina (F) son los más abundantes. Por el contrario el aminoácido hidrofóbico cisteína (C) junto a los aromáticos triptófano (W) y metionina (M) son los menos abundantes. Los 25 genes para el ARN de transferencia identificados no son suficientes para codificar todos los codones puesto que falta el codón para la prolina (P). Además, los aminoácidos más expresados en los genes codificantes mitocondriales leucina e isoleucina, junto a la arginina, glicina y serina, son codificados por dos codones diferentes cada uno.

*Comparación del genoma mitocondrial de Trebouxia TR9 con el de otras algas verdes.*

Si bien la secuencia del genoma mitocondrial de *Trebouxia* sp. TR9 es la única secuencia completa de un genoma mitocondrial de un alga liquénica del género *Trebouxia*, hay otro alga liquénica del mismo género, *Trebouxia aggregata* SAG 219-1d, de la que hay disponibles varias secuencias parciales de su genoma mitocondrial depositadas en GenBank (accesos EU123944 - EU123949). El contenido en genes de estas secuencias de *T. aggregata* es idéntico al del genoma completo de *Trebouxia* sp. TR9 (Tabla 9). Con ello podemos afirmar que *T. aggregata* posee todos los genes identificados en *Trebouxia* sp. TR9. Lo que desconocemos es si *T. aggregata* posee alguno adicional, ya que el genoma de este alga no está completo. Después de analizar el orden de los genes en los diferentes fragmentos del genoma mitocondrial de *T. aggregata* y los homólogos de *Trebouxia* sp. TR9 con el programa

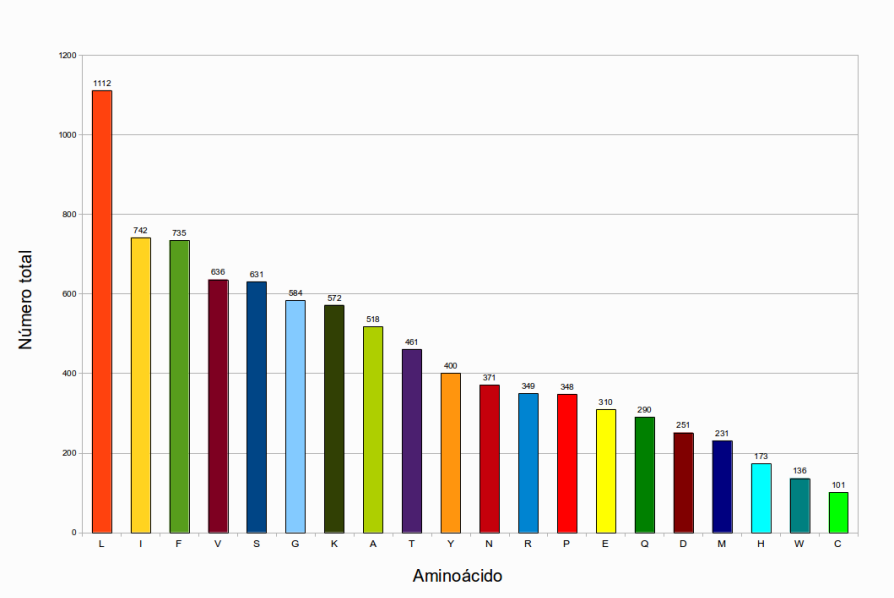



Figura 22: **Número total de cada aminoácido codificado en el genoma mitocondrial de *Trebouxia* sp. TR9.** Alanina (A), Cisteína (C), Ácido aspártico (D), Ácido glutámico (E), Fenilalanina (F), Glicina (G), Histidina (H), Isoleucina (I), Lisina (K), Leucina (L), Metionina (M), Asparagina (N), Prolina (P), Glutamina (Q), Arginina (R), Serina (S), Treonina (T), Valina (V), Triptófano (W), Tirosina (Y).

MAUVE (Darling *et al.* , 2004) (Figura23), una de las características más relevantes es el alto nivel de conservación en el orden de genes entre ambas, en todos los fragmentos analizados. La sintenia es muy elevada con la excepción de cinco genes: *nad9*, *nad3* - *nad7* y *nad6* - *nad4L* que, a diferencia de *T. aggregata*, en *Trebouxia* sp. TR9 se ubican a ambos lados del gen *cob* (Figura24).

Una diferencia notable en la estructura del genoma mitocondrial de ambas algas, es la reducción de las regiones intergénicas en *Trebouxia* sp. TR9 respecto de *T. aggregata*, que presenta unas regiones intergénicas de mayor tamaño. Este hecho explica que los mismos genes estén contenidos en 70.070 nt y al menos 130.056 nt en *Trebouxia* sp. TR9 y *T. aggregata*, respectivamente. Esta característica indica una mayor expansión general del genoma mitocondrial de *T. aggregata* respecto al del de *Trebouxia* sp. TR9. Esta característica es compartida por el resto de algas de la clase Trebouxiophyceae. En la Figura 25 puede observarse el incremento del tamaño de algunos genomas mitocondriales como consecuencia de la existencia de largas regiones intergénicas. El ejemplo más notable es el de *Chlorokybus atmophyticus* (Streptophyta).

Respecto a la presencia de intrones, ambas especies de *Trebouxia* presentan intrones en los mismos genes (*rrnL*, *cox1* y *cob*) con excepción del gen *rrnS* que presenta un intrón en *Trebouxia* sp. TR9 pero no en *T. aggregata*. Al igual que en *Trebouxia* sp. TR9 en *T. aggregata*, los intrones de los genes cox1 y *cob* codifican ORFs que corresponden

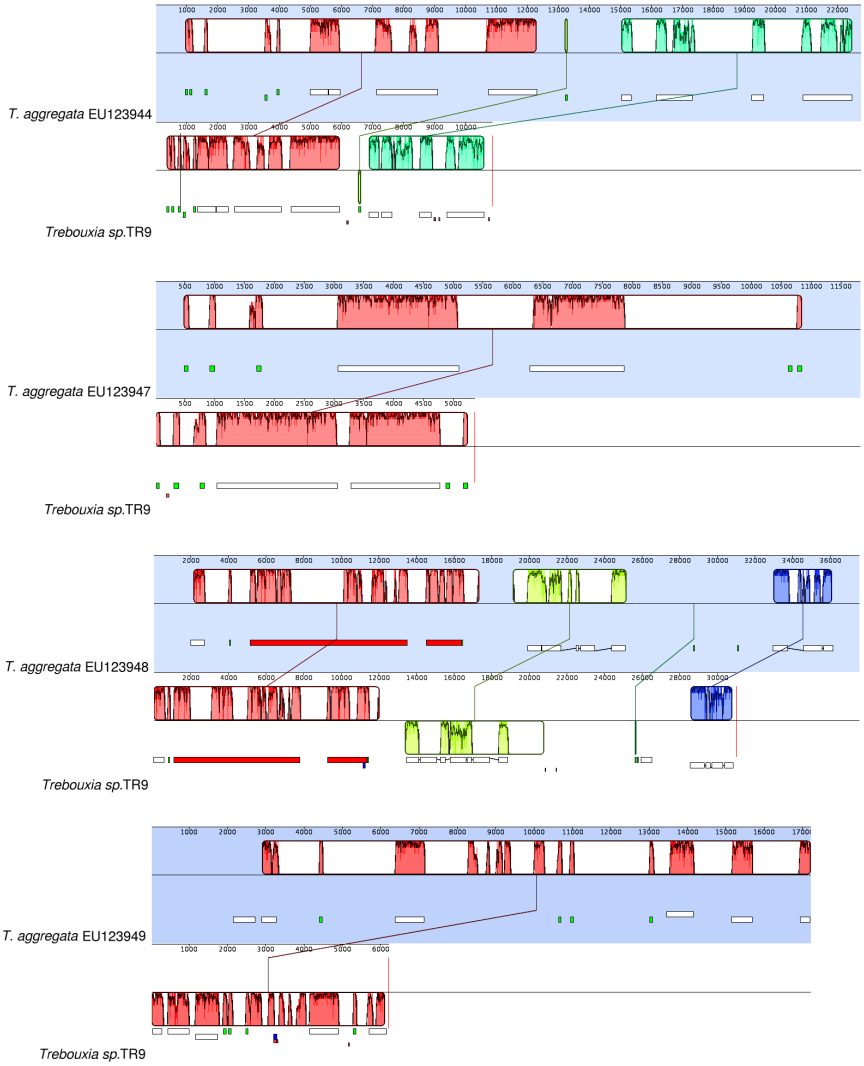



Figura 23: **Alineamiento de las porciones de los genomas mitocondriales secuenciados de *Trebouxia aggregata* y *Trebouxia* sp. TR9.** Los bloques coloreados indican regiones de secuencias genómicas que alinean con partes de otros genomas. Las líneas de colores unen bloques con similitud entre dos genomas.

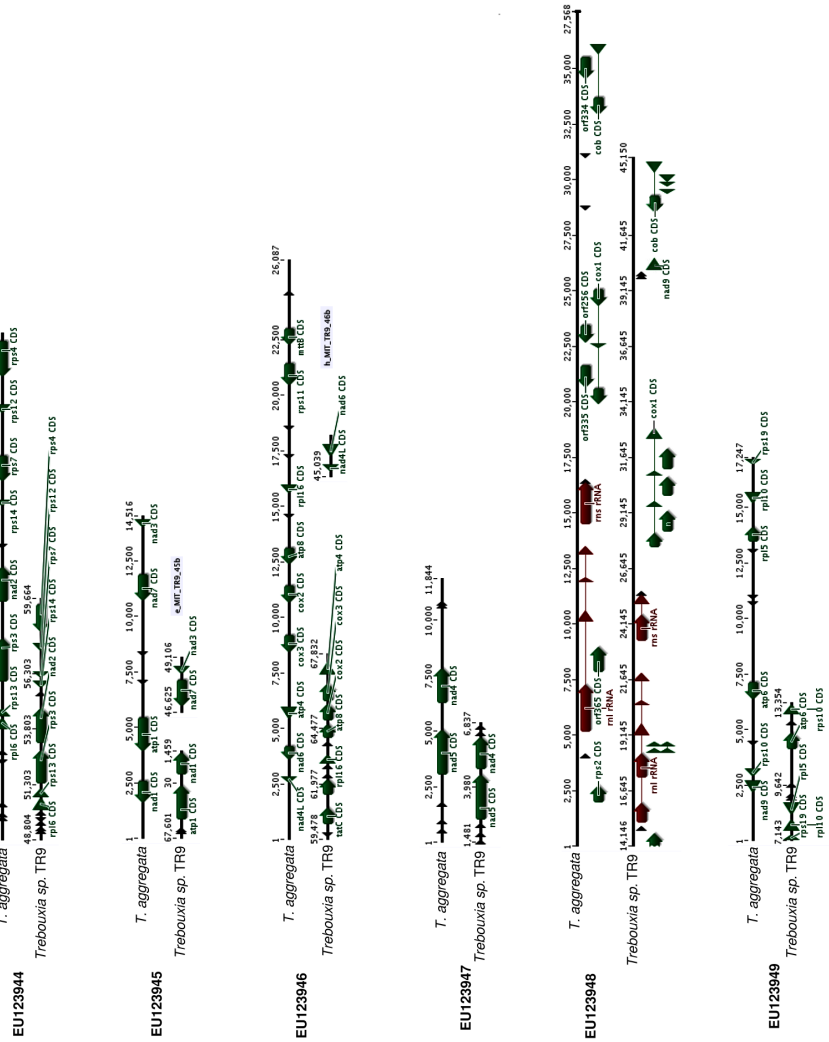



Figura 24: **Mapas genéticos de las regiones de los genomas mitocondriales de *Trebouxia aggregata* y *Trebouxia* sp. TR9.** Las partes del genoma que codifican proteína se señalan en verde, las que codifican ARNr se señalan en granate y las que codifican ARNt se señalan en negro

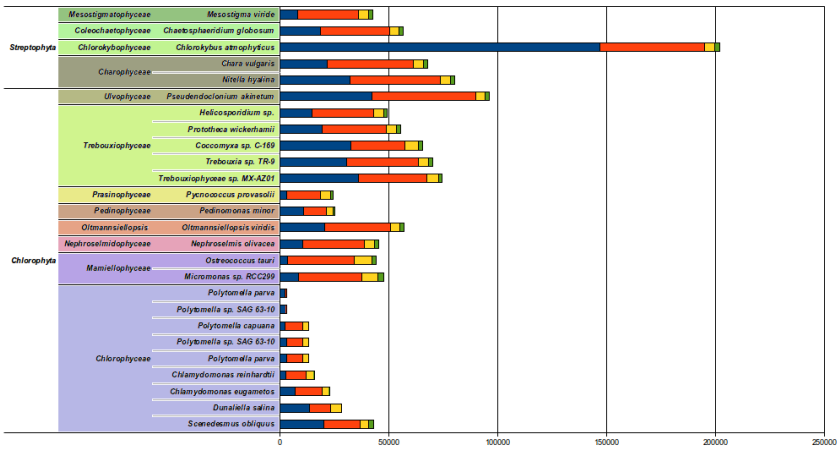



Figura 25: **Proporción de nucleótidos no codificantes, CDS, ARNr y ARNt de genomas mitocondriales de algas verdes de la división Chlorophyta y algunas de la división Streptophyta.** En azul oscuro se indica la parte no transcrita, en rojo la longitud total de CDSs, en amarillo la longitud total de genes ribosomales y en verde la longitud total de ARNt. Cada clase de algas se indica por un color.

con LHEs. En *T. aggregata* a diferencia de *Trebouxia* sp. TR9, uno de los intrones del gen *rrnL* codifica una ORF. Algunos intrones coinciden en el sitio de inserción pero otros no. El gen *cob* presenta un intrón en ambas especies de *Trebouxia* en el mismo sitio de inserción y codifica una LHE similar a las de algas de la división Streptophyta (Figura 26). Los intrones del gen *cox1* difieren en número, apareciendo tres en *Trebouxia* sp. TR9 y solamente dos en *T. aggregata*, todos ellos con ORFs codificantes de LHEs. En cuanto a las posiciones solamente coincide la del intrón 2 de *Trebouxia* sp. TR9. El gen *rrnL* presenta tres intrones en ambas especies de *Trebouxia* cuyas posiciones coinciden para los intrones 2 y 3. El primer intrón difiere en su posición y en que en *T. aggregata* codifica una LHE.

Al comparar el genoma de *Trebouxia* sp. TR9 con el resto de algas de la clase Trebouxiophyceae, aparecen 11 zonas conservadas (Figura 27). En base al bajo nivel de sintenia observado entre estos bloques, es muy probable que se hayan producido intensos reordenamientos a partir de un "orden de genes ancestral" del ancestro común de este grupo de algas.

El contenido de genes de los genomas mitocondriales de algas verdes varía mucho dependiendo del organismo con el que se compare (Tabla 10). Como era de esperar, los genomas mitocondriales cuyo contenido en genes se asemeja más al de *Trebouxia* sp. TR9 son los de las algas de la Clase Trebouxiophyceae. En un extremo, se encuentra el alga no fotosintética *Helicosporidium* sp. que es la especie con la que comparte todos los genes codificantes para proteínas excepto la pro-

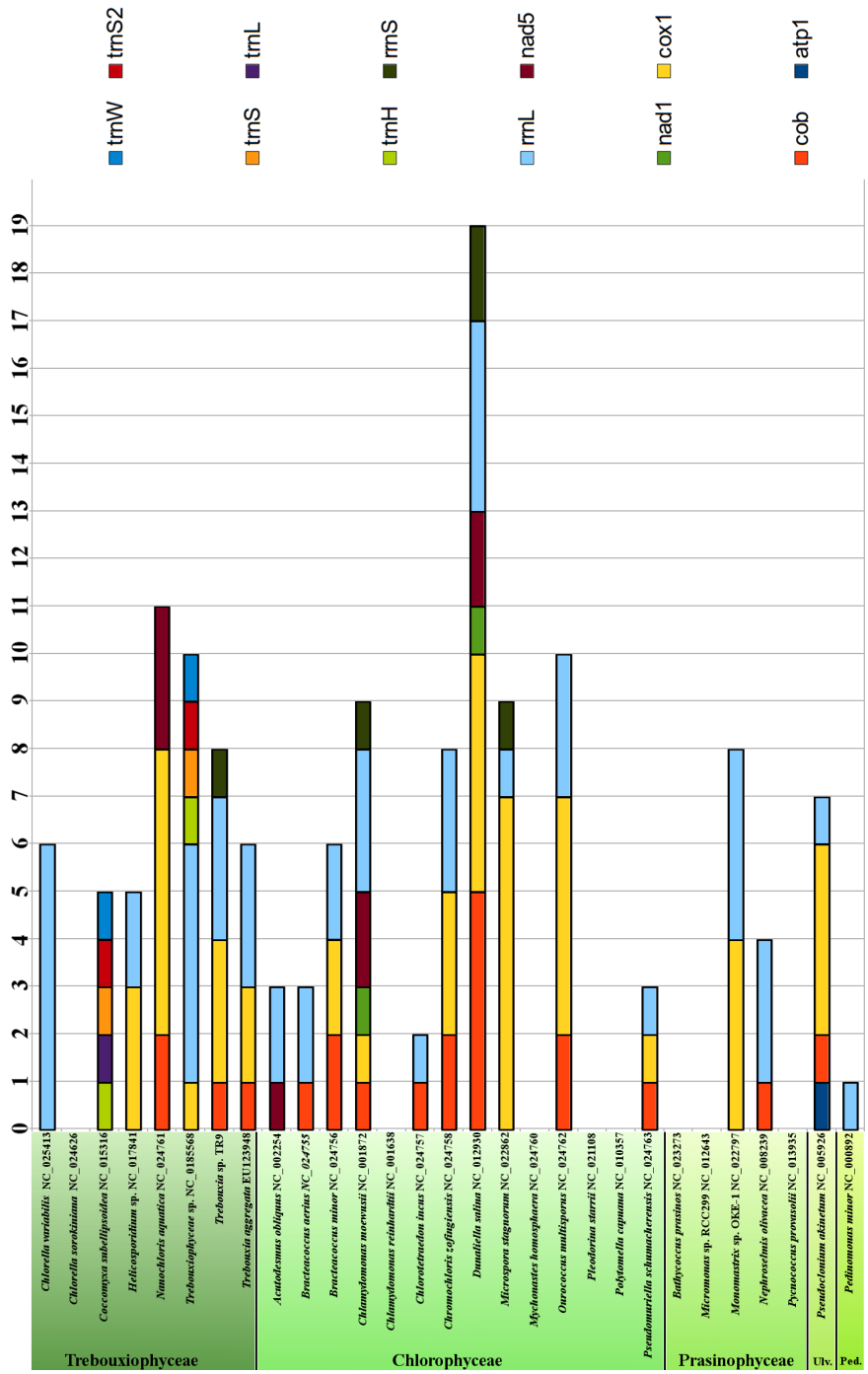



Figura 26: **Distribución y número de intrones en los genes codificados en los genomas mitocondriales de algas verdes de la división Chlorophyta.** Ulvophyceae=Ulv, Pedinophyceae=Ped.

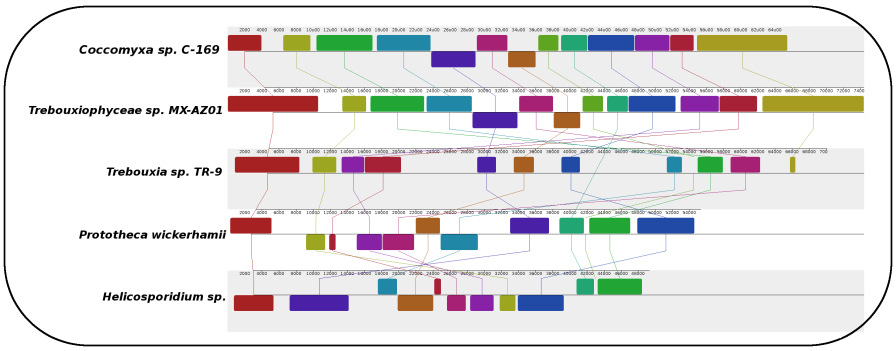
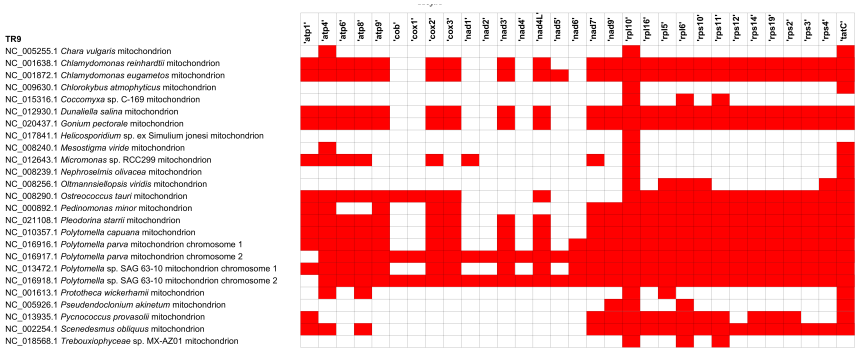



Figura 27: **Alineamiento múltiple de los genomas mitocondriales secuenciados de Trebouxiophyceae.** Los bloques coloreados indican regiones de secuencias genómicas que se alinean con partes de otros genomas. Las líneas de colores unen bloques con similitud entre dos genomas.

Cuadro 10: **Genes codificados en el genoma de *Trebouxia* sp. TR9 presentes (blanco) o ausentes (rojo) en otros genomas mitocondriales de algas verdes.**




teína ribosomal *rpl10*. En el otro extremo, se encuentran las algas no fotosintéticas del género *Polytomella* en cuyos genomas mitocondriales, muy reducidos y lineales, se encuentran ausentes la mayor parte de los genes presentes en *Trebouxia* sp. TR9. También puede observarse una situación similar a la de *Trebouxia* sp. TR9 con respecto a *T. aggregata* entre *Coccomyxa* sp. C-169 y Trebouxiophyceae sp. MX-AZ01 que presentan un alto nivel de conservación del orden de los genes al compartir 13 bloques similares entre ambas. Esto sugiere que, al igual que sucede con *Trebouxia* sp. TR9 respecto de *T. aggregata*, difieren evolutivamente muy poco entre sí y seguramente pertenezcan al mismo género *Coccomyxa*.

En la figura 28 se muestran gráficas de alineamientos de tipo "dotplot". En ella puede observarse como a medida que el genoma presenta un mayor tamaño, aparecen más puntos de similitud azarosa dentro de las regiones intergénicas: desde prácticamente ausentes, como en el caso de *Chlamydomonas reindhartii*, a una gran proporción de regiones intergénicas en *Nicotiana tabacum*. En el caso de *Trebouxia*

sp. TR9, se pueden apreciar unas zonas únicas correspondientes a los genes ribosomales *rnnL* y *rnnS* en torno a las posiciones 15.000 - 20.000 y 23.000 - 25.00 respectivamente. En las posiciones 47.000 - 48.000, 65.000 - 66.000 y 68.000 - 70.000 se encuentran los genes *nad7*, *cox3* y *atp1* con un patrón único similar a los genes ribosomales (Figura 28). Las algas *Ostreococcus* y *Micromonas* también contienen unas regiones repetidas invertidas propias, por el momento, de la clase Mamiellophyceae a la que pertenecen. En la clase Trebouxiophyceae, puede observarse un patrón similar en el genoma mitocondrial de *Trebouxia* sp. TR9, con *Coccomyxa* y Trebouxiophyceae sp. MX-AZ01, en comparación con el alga parásita *Helicosporidium* sp.

*Análisis filogenéticos basados en genes codificados en la mitocondria.*

Con las secuencias del genoma mitocondrial de *Trebouxia* sp. TR9 y las de otras algas verdes disponibles en GenBank, realizamos análisis filogenéticos con el objeto de tener una visión general de la posición de *Trebouxia* sp. TR9 en relación con otras algas más o menos próximas desde un punto de vista evolutivo. En primer lugar, se analizaron los datos bajo hipótesis filogenética de máxima verosimilitud. Se utilizaron secuencias de aminoácidos deducidas las ORFs de los genomas mitocondriales de algas verdes de la división Chlorophyta representativas de las diferentes clases: Chlorophyceae, Mamiellophyceae, Prasinophyceae, Trebouxiophyceae y Ulvophyceae. También se incluyeron dos especies de la división Streptophyta que se utilizaron como grupo externo (*Chara vulgaris* y *Chlorokybus atmophyticus*). Se utilizaron secuencias de aminoácidos ya que al analizar relaciones filogenéticas de divergencias ancestrales muy separadas, los datos basados en aminoácidos son menos propensos a problemas de saturación, de sesgos debidos a convergencia por substituciones sinónimas en el tercer codón o por convergencia en el uso de codones (Li *et al.* , 2014). El estudio filogenómico realizado, se llevó a cabo con un conjunto de proteínas codificadas por estos siete genes: *cob*, *cox1*, *nad1*, *nad2*, *nad5*, *nad6,* que fueron los únicos que estaban presentes en las 25 especies de algas incluidas en el análisis. El árbol filogenético obtenido bajo la hipótesis de máxima parsimonia presentó un índice de consistencia de 0.6589 y un índice de "homoplasia" de 0.3411. El mejor modelo evolutivo se determinó con ayuda del programa ProtTest (Darriba *et al.* , 2011) bajo el Criterio de Información de Akaike (AIC) que resultó ser el modelo LG (Le & Gascuel, 2008) (Tabla 11).

En la Figura 29 se muestra el árbol filogenético resultante del análisis basado en la hipótesis de máxima verosimilitud cuya topología resultó prácticamente idéntica al obtenido con el criterio de máxima parsimonia. Como puede observarse en la Figura 29, los valores de soporte de todas las ramas en general son bastante elevados en los tipos de análisis filogenéticos llevados a cabo. *Trebouxia* sp. TR9 aparece englobada en el clado que recoge todas las algas de la clase Tre-

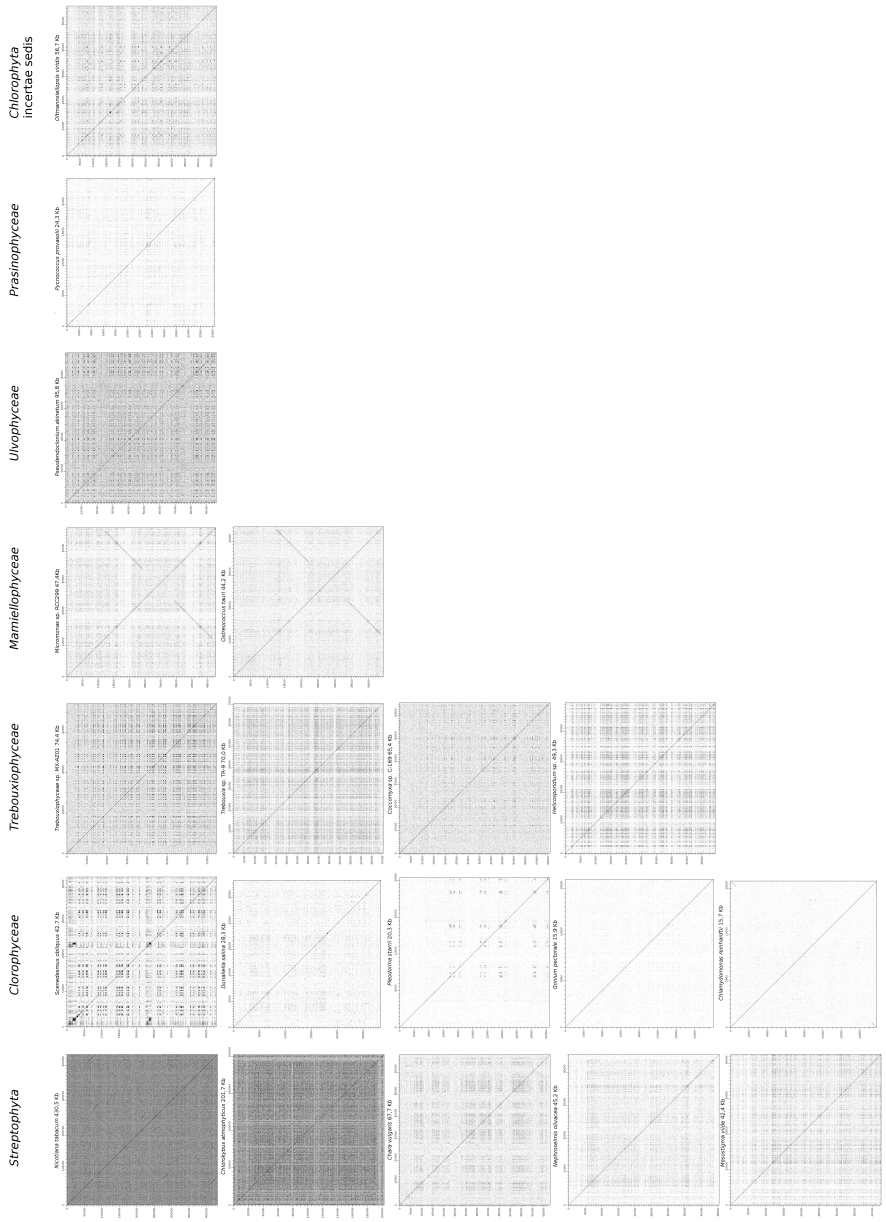



Figura 28: **Alineamiento "dot plot" de genomas mitocondriales de algas verdes. Cada genoma se comparó con sigo mismo utilizando el programa dotter.** Todas las posiciones de la secuencia de cada genoma se comparan con todas las posiciones de esa misma secuencia corriendo una ventana. Dondequiera que aparezca similitud por encima de un límite entre dos posiciones de cada secuencia, se ha sido dibujado un punto.

Cuadro 11: **Resultados de la evaluación de diferentes modelos evolutivos por el programa ProtTest bajo el criterio de Akaike (AIC).**

| Model | deltaAIC* | AIC | -lnL |
|---|---|---|---|
| **LG** | 0,00 | 66889,39 | -33397,69 |
| **Blosum62** | 352,05 | 67241,44 | -33573,72 |
| **CpREV** | 510,79 | 67400,18 | -33653,09 |
| **MtArt** | 781,17 | 67670,55 | -33788,28 |
| **VT** | 850,87 | 67740,26 | -33823,13 |
| **WAG** | 899,00 | 67788,39 | -33847,19 |
| **RtREV** | 1157,21 | 68046,59 | -33976,30 |
| **JTT** | 1391,62 | 68281,01 | -34093,51 |
| **MtREV** | 1671,54 | 68560,92 | -34233,46 |
| **DCMut** | 2531,44 | 69420,83 | -34663,42 |
| **Dayhoff** | 2538,64 | 69428,03 | -34667,01 |
| **MtMam** | 3754,46 | 70643,85 | -35274,93 |
| **HIVb** | 4027,38 | 70916,76 | -35411,38 |
| **HIVw** | 7767,79 | 74657,18 | -37281,59 |

bouxiophyceae muy próxima a otro alga del mismo género *Trebouxia aggregata*, también liquénica. Junto a ellas se encuentran dos especies del género *Coccomyxa* formando las cuatro especies un clado con un elevado soporte (98/97). Más alejado se encuentra otro clado con elevados valores de soporte (99/100) que incluye el resto de algas de la clase Trebouxiophyceae y cuyas ramas presentan también elevados valores de credibilidad (>99). El clado que incluye las algas de la clase Trebouxiophyceae (T) forma, junto con las clases Ulvophyceae (U) y Chlorophyceae (C), el clado UTC con un soporte moderado (66/82). Las especies de la clase Prasinophyceae aparecen en un clado más externo relacionado con el clado UTC. Especies de clases definidas más recientemente como Mamiellophyceae o Pedinophyceae, aparecen dentro del clado UTC. Ambas clases parecen más próximas a las clases Chlorophyceae y Ulvophyceae que a la clase Trebouxiophyceae. Con tan solo siete genes, hemos logrado una filogenia bastante bien resuelta con valores de soporte bastante aceptables.

### 4.2.2 *Discusión*

Los genomas organulares, especialmente los de mitocondrias, son todavía poco conocidos en algunos grupos de organismos como por ejemplo las algas de la división Chlorophyta, de los que hay completamente secuenciados apenas una treintena frente a los cerca de 60 genomas de cloroplastos de organismos de esta división y unos 80

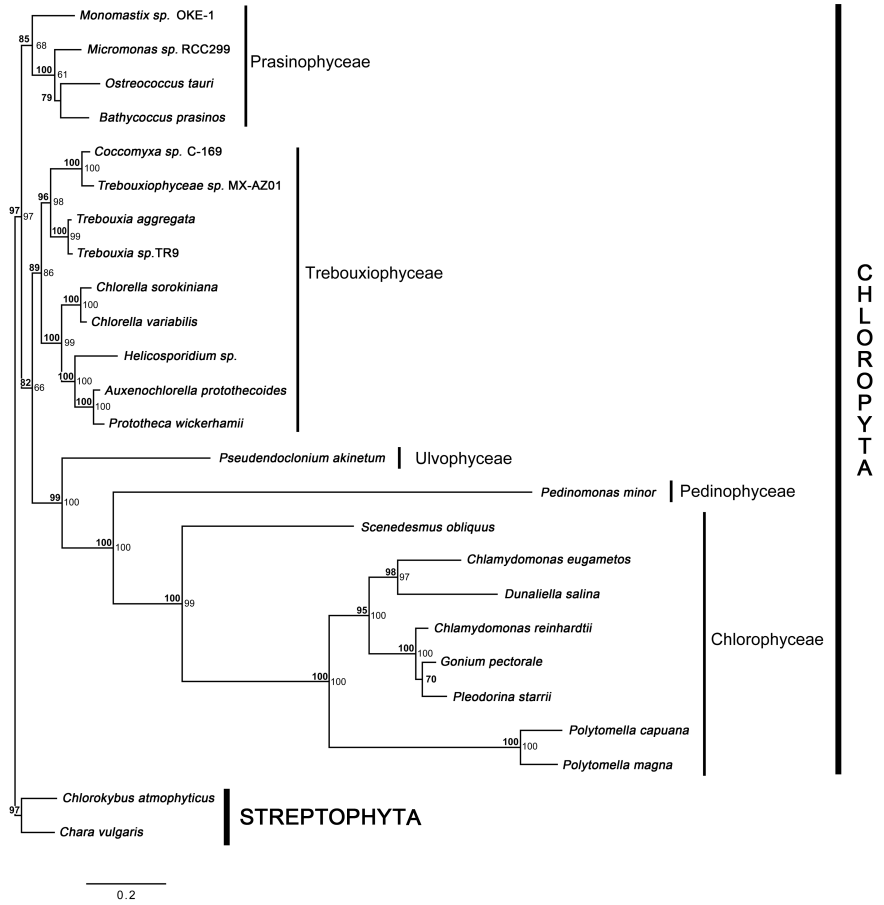



Figura 29: **Filogenia basada la secuencia de aminoácidos de siete proteínas mitocondriales codificadas por los genes *cob*, *cox1*, *nad1*, *nad2*, *nad5* y *nad6*.** Los números sobre cada nodo muestran los valores de probabilidades posteriores derivados del análisis de máxima parsimonia y los situados a la derecha muestran los valores de "bootstrap" derivados del análisis de máxima verosimilitud. La barra de escala representa las distancias genéticas en sustituciones de aminoácidos por posición.

genomas mitocondriales de organismos de la división Streptophyta. Los genomas organulares de las plantas terrestres presentan una gran variedad de tamaños, transferencias horizontales de genes y frecuentes reordenamientos (e.g., Palmer & Herbon 1988; Adams *et al.* 2002; Mach 2011). Dada la enorme variación en tamaño, estructura y contenido de los genomas mitocondriales de algas, algunos de los cuales utilizan códigos genéticos diferentes como es el caso de *Scenedesmus obliquus* (Nedelcu *et al.* , 2000), es de esperar que encontremos una mayor diversidad si estudiamos más genomas mitocondriales dentro de grupos de algas específicos que pueden definirse con diferentes criterios (evolutivo, ecológico. . . ). Siguiendo un criterio evolutivo, un estudio reciente de varios genomas mitocondriales de algas verdes del orden *Sphaeropleales* ha mostrado una diversidad considerable y una serie de características peculiares en comparación con otras algas verdes (Fučíková *et al.* , 2014). Dentro de la división Chlorophyta, la clase Trebouxiophyceae incluye numerosos y diversos organismos tanto simbiontes mutualistas (por ejemplo las algas liquénicas) como parásitas, saprofitas y de vida libre. Un estudio reciente (Lemieux *et al.* , 2014), aporta 27 secuencias de genomas de cloroplastos y establece posibles relaciones filogenéticas dentro de esta clase y en relación con las demás clases de la división Chlorophyta. La mayoría de las algas liquénicas pertenecen al género *Trebouxia,* sin embargo, no se dispone de la secuencia completa del genoma mitocondrial de ninguna especie de este género. Únicamente contamos con secuencias parciales de *Trebouxia aggregata* (accesos EU123944-EU123949). En la presente Tesis Doctoral se ha obtenido la primera secuencia completa del genoma mitocondrial de un alga del género *Trebouxia.*

*Tamaño del genoma, ordenación y genes codificados*

En el genoma mitocondrial de *Trebouxia* sp. TR9 tiene un tamaño 70.070 nt que es uno de los mayores si lo comparamos con otros genomas mitocondriales secuenciados en organismos de la división Chlorophyta cuyo tamaño promedio se sitúa en aproximadamente 44.438 nt (a 9 de marzo de 2015). Si se ordenan los tamaños de 38 genomas mitocondriales secuenciados en algas de esta división (disponibles en el NCBI a fecha de 10 de marzo de 2015), el de *Trebouxia* sp. TR9 ocupa el cuarto puesto, solamente por detrás de *Pseudendoclonium akinetum* (9.5880 nt), *Microspora stagnorum* (87.405 nt), *Chlorella variabilis* (78.500) y Trebouxiophyceae sp. MX-AZ01 (744.23 nt). En general, puede afirmarse que, hasta la fecha, dentro de los genomas mitocondriales de las algas de la división Chlorophyta, los de algas de la clase Trebouxiophyceae son los que presentan en general un mayor tamaño, solamente superado por *Pseudendoclonium akinetum*, único representante secuenciado de la clase Ulvophyceae. En este sentido, sorprende el tamaño del genoma de *Trebouxia aggregata* que se estima ser superior a 130.000 nt.

Los genomas mitocondriales de plantas terrestres presentan un tamaño promedio del genoma mitocondrial mucho mayor que las algas verdes (aproximadamente 540.890 nt) con una extraordinaria variabilidad desde 11.318.806 nt en la planta terrestre *Silene conica* **(Sloan *et al.* , 2010)** hasta alrededor de 100.000 nt en los musgos **(Liu *et al.* , 2014)**. En la división Streptophyta, los genomas mitocondriales de menor tamaño corresponden a algas como *Nitella hialina*, *Roya obtusa*, *Chara vulgaris*, *Entransia fimbriata*, *Chaetosphaeridium globosum* y *Mesostigma viride* con tamaños de 80.193 nt, 69.465 nt, 67.737 nt, 61.645 nt, 56.574 nt y 42.424 nt, respectivamente, similares a los de las algas de la división Chlorophyta, seguidos por los de varias plantas no vasculares -musgos- con tamaños de aproximadamente unos 100.000 nt (Liu *et al.* , 2014). Hay excepciones a esta regla, como por ejemplo *Chlorokybus atmophyticus* **(Turmel *et al.* , 2007),** que a pesar de ser un alga unicelular, tiene un genoma mitocondrial de 201.763 nt. El gran tamaño del genoma de este organismo se debe, más que a la presencia de un número sensiblemente superior de genes codificantes de proteínas, a que tiene regiones intergénicas más largas. Respecto a la proporción de regiones no codificantes en comparación con las codificantes, *Trebouxia* sp. TR9 presenta una proporción ligeramente superior al resto de las algas de la división Chlorophyta. Como ya se ha indicado anteriormente, esta característica es más acusada para *Trebouxia aggregata* y es compartida por otras algas de la clase Trebouxiophyceae.

*Presencia y diversidad de intrones*

En *Trebouxia* sp. TR9 hemos encontrado un total de 9 intrones distribuidos en cuatro genes: *cob*, *cox1, rrnL* y *rrnS*. Si analizamos las secuencias de los genomas mitocondriales de las algas de la división Chlorophyta disponibles en GenBank, puede observarse como el número de intrones es muy variable, desde 19 en *Dunaliella salina* NC_012930 (Smith *et al.* , 2010) hasta ninguno en *Chlamydomonas reinhardtii* NC_ 001638 (Vahrenholz *et al.* , 1993), ambas de la clase Chlorophyceae, *Micromonas* sp. NC_ 012643 (Worden *et al.* , 2009) de la clase Prasinophyceae y *Chlorella sorokiniana* NC_024626 (Orsini *et al.* , 2014) de la clase Trebouxiophyceae. El promedio de número de intrones por genoma mitocondrial en las algas de la división Chlorophyta disponibles hasta la fecha (Abril de 2015), se sitúa aproximadamente en unos seis (Tabla 3.5). Todos los intrones encontrados en *Trebouxia* sp. TR9 pertenecen al grupo I (IA, IB y ID), que son los más abundantes en los genomas mitocondriales de la mayoría de las algas de la división Chlorophyta. La presencia de intrones no parece aleatoria, concentrándose en una serie de 12 genes de los 29 genomas mitocondriales de algas analizados de la división Chlorophyta. Además, hay ciertos genes en los que hay una tendencia general a la concentración de intrones, como por ejemplo los genes *rrnL*, *cob* y *cox1* que suman

un total de 50, 47 y 21 intrones en los 29 genomas analizados. Estos genes son precisamente los únicos que contienen intrones en *Trebouxia* sp. TR9. En las algas de Chlorophyta, el número de intrones en el genoma mitocondrial es mucho más bajo que en la división Streptophyta. En esta última, encontramos organismos que tienen cerca de la veintena de intrones como la planta *Arabidopsis thaliana* (Unseld *et al.* , 1997), con 21 intrones y el alga *Chlorokybus atmophyticus* (Turmel *et al.* , 2007) con 20 intrones. Pero también encontramos casos con un número menor como *Mesostigma viride (Turmel* et al. , *2002)* con 9 intrones, que se acerca más a los valores de las algas de la división Chlorophyta. Los genes que contienen intrones en los genomas mitocondriales en Streptophyta son más variables y diferentes de los de Chlorophyta, siendo muy abundante su presencia en genes que codifican subunidades de la cadena respiratoria mitocondrial (p.e. 18 intrones de los 20 presentes en el genoma mitocondrial de *Arabidopsis thaliana*).

Otro tipo de intrones muy frecuentes en genomas organulares, tanto de mitocondrias como de cloroplastos son los del grupo II. Como se ha mencionado anteriormente, en el genoma mitocondrial de *Trebouxia* sp. TR9 no se han identificado intrones de tipo II enteros, que sí están presentes en los genomas mitocondriales de otras algas de la división Chlorophyta, incluso en las pertenecientes a la clase Trebouxiophyceae, como por ejemplo *Coccomyxa subellipsoidea* C-169 NC_015316 (Smith *et al.* , 2011). Sin embargo, en el genoma mitocondrial de *Trebouxia* sp. TR9 y *Trebouxia aggregata* (EU123949.1) se ha identificado una pequeña secuencia de 75 nt entre los genes *trnG* (UCC) y *atp6* correspondiente a los dominios V y VI propios de intrones de tipo II. Una posible interpretación de la presencia de esta secuencia es que se trate probablemente de una reminiscencia de un intrón del grupo II preexistente en un linaje ancestral que se perdió a lo largo de la evolución con el gen que lo contenía. Búsquedas de secuencias similares en los genomas mitocondriales de algas de la división Chlorophyta, mostraron la presencia de motivos similares en el extremo 3' de varios intrones completos del tipo II. Algunos ejemplos son: *Scenedesmus obliquus* (NC 002254) (Kück *et al.* , 2000), *Bracteacoccus aerius* (NC_024755), *Bracteacoccus minor* (NC_024756) y *Chromochloris zofingiensis* (NC_024758) (Fučíková *et al.* , 2014) de la clase Chlorophyceae en el gen *rrnL*; *Monomastix* sp. OKE-1 (NC_022797) (Turmel *et al.* , 2013) de la clase Prasinophyceae, *Coccomyxa subellipsoidea* C-169 (NC_015316) (Smith *et al.* , 2011) y Trebouxiophyceae sp. MX-AZ01 (NC_018568) (Servín-Garcidueñas & Martínez-Romero, 2012) de la clase Trebouxiophyceae en el gen *trnS*. En todos los casos, las secuencias referidas, forman parte de intrones de tipo II completos excepto en el caso de *Monomastix* sp. en el que, al igual que en *Trebouxia* sp. TR9 y *T. aggregata*, no forma parte de un intrón de tipo II completo.

*Preferencia de codones y código genético*

La secuenciación masiva de genomas ha aumentado la disponibilidad de secuencias de nucleótidos de especies poco estudiadas con códigos genéticos diferentes de los conocidos. Un buen ejemplo es el genoma mitocondrial del alga verde *Scenedesmus obliquus* (Chlorophyta) (Nedelcu *et al.* , 2000). Ello llevó a proponer nuevos códigos genéticos como el de mitocondrias de "Chlorophyceae" (NCBI translation table 16) y el de *"Scenedesmus opliquus"* (NCBI translation table 22). Actualmente, aproximadamente el 0,65 % de las secuencias de ADN depositadas en GenBank codifican su proteína con un código genético que se desvía de los conocidos hasta la fecha (Dutilh *et al.* , 2011). Cuando se trata de traducir una secuencia de nucleótidos que se supone que codifica alguna proteína, si el organismo al que pertenece utiliza un código diferente del "estándar", la predicción de ORFs en base a este código genético, puede generar serios errores con consecuencias importantes, tanto para análisis filogenéticos como para estudios orientados a la funcionalidad de las proteínas. Por este motivo, se debe determinar qué tipo de código genético utiliza el genoma analizado del organismo en cuestión. A la hora de utilizar programas informáticos que facilitan la labor de determinar el código genético en cuestión tales como Gendecoder (Abascal *et al.* , 2006) o FACIL (Dutilh *et al.* , 2011) es muy importante estar seguros de que el conjunto de secuencias objeto de análisis pertenecen al mismo sistema (núcleo, cloroplasto o mitocondria). En nuestro caso, la obtención del genoma completo circular de la mitocondria de *Trebouxia* sp. TR9 asegura la pertenencia de toda la secuencia sometida al análisis al mismo genoma. Con se ha indicado en la parte de resultados, el programa FACIL predijo el código genético estándar como el más probable para el genoma mitocondrial de *Trebouxia* sp. TR9. Consecuentemente, todas las ORFs y secuencias de aminoácidos se dedujeron en base a este código. Es importante señalar que el genoma mitocondrial de *Neochloris aquatica* NC_024761, que es un alga verde de la clase Trebouxiophyceae al igual que *Trebouxia* sp. TR9, presenta el código genético de *"Scenedesmus obliquus"* (acceso NC_02476; NCBI translation table=22) en lugar del "estándar".

*Análisis filogenéticos basados en secuencias de genes mitocondriales*

La filogenia representada en la Figura 29 que incluye diversas algas de diferentes clases de la división Chlorophyta, presenta valores de soporte bastante aceptables en casi todos los nodos. Con la utilización de las secuencias de tan solo siete genes (*cob*, *cox1*, *nad1*, *nad2*, *nad4*, *nad5* y *nad6*), los únicos compartidos por las algas incluidas en la filogenia de la Figura 29, se ha logrado una filogenia bastante bien resuelta y robusta. Este hecho sugiere que los genes de mitocondrias pueden ser bastante prometedores para estudiar las relaciones filoge-

néticas de algas verdes tal y como adelantan Tippery *et al.* (2012) en su trabajo sobre el orden *Sphaeropleales*, aunque en este caso utilizaron cuatro genes cloroplásticos (*psaA*, *psaB*, *psbC*, *rbcL*) y un gen ribosomal nuclear (18S). Las relaciones entre las diferentes clases en la filogenia de la Figura 29 presentan analogías y diferencias con una filogenia recientemente publicada basada en genes cloroplásticos (Lemieux *et al.* , 2014). En líneas generales, nuestra filogenia coincide con la de Lemieux *et al.* (2014) en la posición relativa de las clases Prasinophyceae, Chlorophyceae, Trebouxiophyceae y Ulvophyceae. Sin embargo, existen dos diferencias fundamentales: (i) la monofilia de la clase Trebouxiophyceae y (ii) la posición de la clase Pedinophyceae. Respecto a la clase Trebouxiophyceae, en nuestra filogenia es un grupo monofilético ya que todas las algas de esta clase ("core Trebouxiophyceae" + *Chlorellales*) aparecen dentro de un mismo clado mientras que en la filogenia de Lemieux *et al.* (2014) *Chlorellales* aparece como un "clado hermano" respecto de las "core Trebouxiophyceae". Los valores de credibilidad para la monofilia de la clase Trebouxiophyceae son de 89/86 (este estudio) y para la parafilia de "core Trebouxiophyceae" y *Chlorellales* es de 88/85 (Lemieux *et al.* , 2014). Probablemente, serán necesarios análisis adicionales para resolver esta duda. En lo referente a la posición de la clase Pedinophyceae, en nuestra filogenia el único representante de este grupo incluido en el análisis (*Pedinomonas minor*) aparece claramente relacionado con la clase Chlorophyceae, con valores de credibilidad de 100/100, mientras que en la filogenia de Lemieux *et al.* (2014) aparece más relacionado con *Chlorellales*, con valores de credibilidad bastante inferiores a los nuestros (84/66). Este hecho nos lleva a pensar que quizá nuestra filogenia se aproxime más a la realidad.

## 4.3 GENOMA CLOROPLÁSTICO

### 4.3.1 *Resultados*

Se ha obtenido la secuencia del genoma de cloroplastos de *Trebouxia* sp. TR9. Para conseguir este genoma, se ha utilizado la tecnología de secuenciación masiva ROCHE 454. En el año 2011 se realizó una primera pirosecuenciación ROCHE 454 GS FLX Titanium (1/4 de placa) y un año más tarde se realizó una segunda pirosecuenciación en la plataforma ROCHE GS JUNIOR en la Unidad de Genómica SCSIE-Universitat de València. Para esta última, se utilizó una "librería paired-end" con un inserto de tamaño aproximado de 3Kb. Dichas secuencias fueron sometidas a un proceso de filtrado por calidad y separación de las lecturas del adaptador de unión ("linker") para los "paired-end", obteniéndose un total de 258.519 lecturas únicas. Todas las lecturas fueron ensambladas "de novo" con el programa GS de Novo Assembler (Newbler) y para su posterior procesado y unión

de "contigs", se mapearon con el software MIRA (Ver más detalles en Materiales y Métodos). Los “scaffolds” pertenecientes al genoma cloroplástico se han filtrado de los de origen nuclear y mitocondrial utilizado el algoritmo de alineamiento local BLAST y en especial los programas BLASTn, BLASTx y tBLASTx contra unas bases de datos de nucleótidos y proteínas de los genomas cloroplásticos secuenciados del NCBI. Del total de “scaffolds” obtenidos, se ha encontrado que tres eran pertenecientes al cloroplasto. Un primer “scaffold” que contenía del "contig" 1 al "contig" 95, el segundo “scaffold” del 97 al 112 y, en el último “scaffold” únicamente el "contig" 278. Para poder discernir las interrelaciones y el orden de los scaffolds cloroplásticos, se ha utilizado el “script bb.contignet”. Dicho “script” filtra la salida del ensamblador Newbler utilizando la información de las lecturas “paired-end” y genera una imagen de dichas interrelaciones (Figura 30). En base a este resultado se han podido unir los tres “scaffolds”. Esta unión se ha comprobado con la ayuda de PCRs en sus extremos, corroborado así la naturaleza circular de este genoma y la existencia de una zona invertida repetida que era el “scaffold” 3 que unía los otros dos “scaffolds”. El "contig" 278 tenía una profundidad 4,6 veces mayor que la media de 60,13 secuencias por nucleótido de los "contigs" pertenecientes a otros “scaffolds” que lo rodeaban (Figura 30). Este dato indica que se corresponde con una zona repetida que había sido colapsada por el ensamblador en un único "contig", lo que corroboraría la existencia de regiones invertidas repetidas en el genoma cloroplástico. PCRs realizadas con los mismos cebadores posicionados en los extremo 5’ o 3’ del "contig" 278, que actuaban bien en sentido directo o inverso dependiendo del cebador complementario que se utilizase en los “scaffolds” 1 o 3 (Tabla 4), corroboraron la naturaleza doble de este "contig".

En la Figura 31 se muestra el resultado final de las uniones de los “scaffolds” 1, 2 y 3 junto a los "contigs" que los formaban. Del total de 113 "contigs" originales ensamblados, con la ayuda de diferentes oligos diseñados y PCRs realizadas, el número de "contigs" se ha disminuido a 49. Con la información de las lecturas ROCHE GS JUNIOR “paired-end”, el ensamblador Newbler ha calculado la distancia entre los "contigs" (Filas con fondo negro de la tabla izquierda de la Figura 31). Estas 49 uniones finales entre "contigs" fueron imposibles de amplificar por PCR debido, en gran medida, a que como se puede apreciar en la Figura 32 el genoma de cloroplastos de *Trebouxia* sp. TR9 contiene muchas repeticiones que han impedido la correcta amplificación de las PCRs y el ensamblaje de estos "contigs".

*Tamaño del genoma, estructura y genes codificados*

El genoma cloroplástico de *Trebouxia* sp. TR9 es de naturaleza circular y presenta un tamaño superior a 300 Kb, Como puede apreciarse en la Figura 32, posee la estructura cuatripartita típica de los cloro-

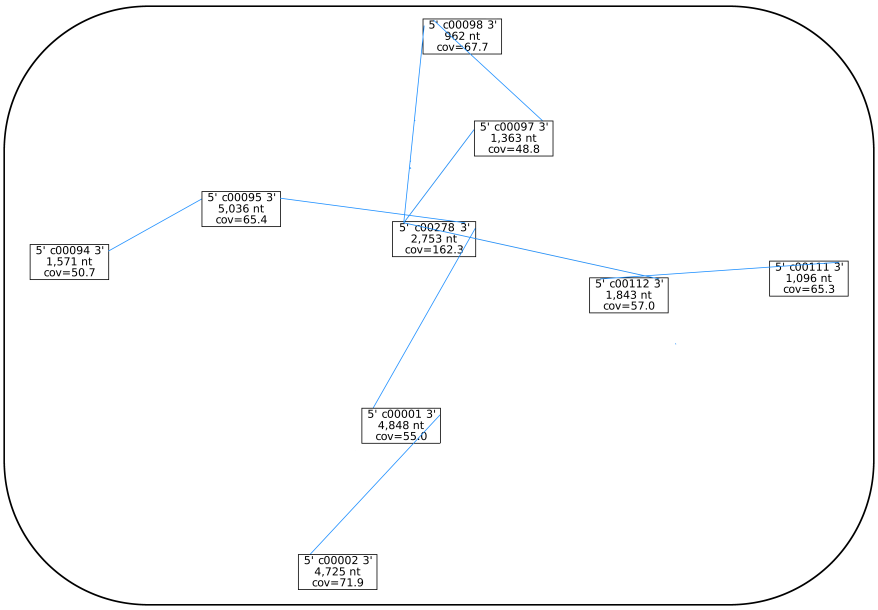
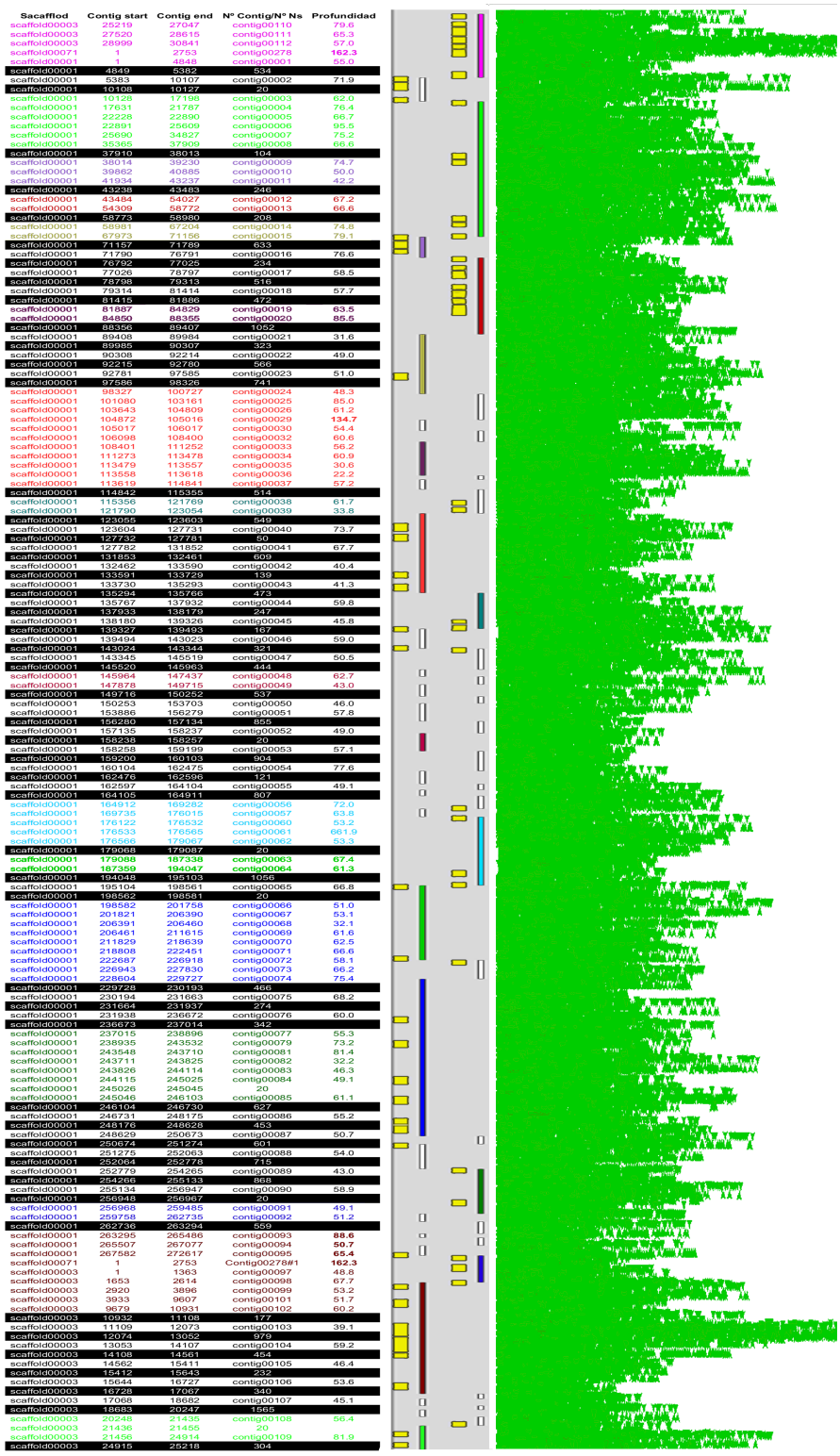



Figura 30: **Esquema de las relaciones entre los "contigs" finales de los "scaffolds" cloroplásticos.** Cada caja representa un "contig", donde se señalan sus extremos 5' y 3', la longitud en nucleótidos (nt) y su profundidad de secuenciación (cov.). Las líneas en azul unen las parejas de lecturas "paired-end" que unen dos "contigs".

plastos de plantas terrestres. Esta estructura consiste en dos regiones repetidas invertidas o IR ("inverted repeats", IRa e IRb) separadas por una región larga de copia única o LSC ("large single-copy") y otra región corta de copia única o SSC ("short single-copy"). En el caso de *Trebouxia* sp. TR9, las regiones IR incluyen un único gen, el *rbcL*, que codifica la sub-unidad grande de la RuBisCO (Figura 32). La proporción de zonas codificantes asciende al 26,73 % del total (303.323 nt). La media de longitud de todos los genes es de 815 nt, mientras que el de los CDSs es de 869 nt, los tRNAs miden de media 70 nt y la media de los genes ribosomales es de 1.390 nt. El % de GC total del genoma completo es del 29,53 % o del 31,92 %, si se cuentan los nucleótidos marcados como "n" en el "scaffold" obtenido o no, respectivamente. El de los genes que codifican proteínas, ARNrs y ARNts presentan un promedio de un 34,2 %, 46,4 % y 53,8 %, respectivamente.

Se han identificado un total de 108 genes en el genoma de *Trebouxia* sp. TR9 (Tabla 12). De todos ellos, 77 codifican proteínas de funciones diversas. Cinco genes codifican proteínas que forman parte de la maquinaria de expresión del genoma cloroplástico: tanto en la transcripción de genes formando parte de la ARN polimerasa (genes *rpoA*, *B*, *C1* y *C2*) como en la traducción de ARNm constituyendo las dos sub-unidades de los ribosomas (genes *rps2*, *3*, *7*, *8*, *9*, *11*, *12*, *14*, *18*, *19* para la sub-unidad pequeña y *rpl2*, *5*, *12*, *14*, *16*, *19*, *20*, *23*, *32*, *36* para la sub-unidad grande) así como en la degradación de pro-



| Sacaflod | Contig start | Contig end | Nº Contig/Nº Ns | Profundidad |
|---|---|---|---|---|
| scaffold00003 | 25219 | 27047 | contig00110 | 79.6 |
| scaffold00003 | 27520 | 28615 | contig00111 | 65.3 |
| scaffold00003 | 28999 | 30841 | contig00112 | 57.0 |
| scaffold00071 | 1 | 2753 | contig00278 | **162.3** |
| scaffold00001 | 1 | 4848 | contig00001 | 55.0 |
| scaffold00001 | 4849 | 5382 | 534 | |
| scaffold00001 | 5383 | 10107 | contig00002 | 71.9 |
| scaffold00001 | 10108 | 10127 | 20 | |
| scaffold00001 | 10128 | 17198 | contig00003 | 62.0 |
| scaffold00001 | 17631 | 21787 | contig00004 | 76.4 |
| scaffold00001 | 22228 | 22890 | contig00005 | 66.7 |
| scaffold00001 | 22891 | 25609 | contig00006 | 95.5 |
| scaffold00001 | 25690 | 34827 | contig00007 | 75.2 |
| scaffold00001 | 35365 | 37909 | contig00008 | 66.6 |
| scaffold00001 | 37910 | 38013 | 104 | |
| scaffold00001 | 38014 | 39230 | contig00009 | 74.7 |
| scaffold00001 | 39862 | 40885 | contig00010 | 50.0 |
| scaffold00001 | 41934 | 43237 | contig00011 | 42.2 |
| scaffold00001 | 43239 | 43483 | 245 | |
| scaffold00001 | 43484 | 54027 | contig00012 | 67.2 |
| scaffold00001 | 54309 | 58772 | contig00013 | 66.6 |
| scaffold00001 | 58773 | 58980 | 208 | |
| scaffold00001 | 58981 | 67204 | contig00014 | 74.8 |
| scaffold00001 | 67973 | 71156 | contig00015 | 79.1 |
| scaffold00001 | 71157 | 71789 | 633 | |
| scaffold00001 | 71790 | 76791 | contig00016 | 76.6 |
| scaffold00001 | 76792 | 77025 | 234 | |
| scaffold00001 | 77026 | 78797 | contig00017 | 58.5 |
| scaffold00001 | 78798 | 79313 | 516 | |
| scaffold00001 | 79314 | 81414 | contig00018 | 57.7 |
| scaffold00001 | 81415 | 81886 | 472 | |
| scaffold00001 | 81887 | 84829 | contig00019 | 63.5 |
| scaffold00001 | 84850 | 88355 | contig00020 | **85.5** |
| scaffold00001 | 88356 | 89407 | 1052 | |
| scaffold00001 | 89408 | 89984 | contig00021 | 31.6 |
| scaffold00001 | 89985 | 90307 | 323 | |
| scaffold00001 | 90308 | 92214 | contig00022 | 49.0 |
| scaffold00001 | 92215 | 92780 | 566 | |
| scaffold00001 | 92781 | 97585 | contig00023 | 51.0 |
| scaffold00001 | 97586 | 98326 | 741 | |
| scaffold00001 | 98327 | 100727 | contig00024 | 48.3 |
| scaffold00001 | 101080 | 103161 | contig00025 | 85.0 |
| scaffold00001 | 103643 | 104809 | contig00026 | 61.2 |
| scaffold00001 | 104872 | 105016 | contig00029 | **134.7** |
| scaffold00001 | 105017 | 106017 | contig00030 | 54.4 |
| scaffold00001 | 106098 | 108400 | contig00032 | 60.6 |
| scaffold00001 | 108401 | 111252 | contig00033 | 56.2 |
| scaffold00001 | 111273 | 113478 | contig00034 | 60.9 |
| scaffold00001 | 113479 | 113557 | contig00035 | 30.6 |
| scaffold00001 | 113558 | 113618 | contig00036 | 22.2 |
| scaffold00001 | 113619 | 114841 | contig00037 | 57.2 |
| scaffold00001 | 114842 | 115355 | 514 | |
| scaffold00001 | 115356 | 121769 | contig00038 | 61.7 |
| scaffold00001 | 121790 | 123054 | contig00039 | 33.8 |
| scaffold00001 | 123055 | 123603 | 549 | |
| scaffold00001 | 123604 | 127731 | contig00040 | 73.7 |
| scaffold00001 | 127732 | 127781 | 50 | |
| scaffold00001 | 127782 | 131852 | contig00041 | 67.7 |
| scaffold00001 | 131853 | 132461 | 609 | |
| scaffold00001 | 132462 | 133590 | contig00042 | 40.4 |
| scaffold00001 | 133591 | 133729 | 139 | |
| scaffold00001 | 133730 | 135293 | contig00043 | 41.3 |
| scaffold00001 | 135294 | 135766 | 473 | |
| scaffold00001 | 135767 | 137932 | contig00044 | 59.8 |
| scaffold00001 | 137933 | 138179 | 247 | |
| scaffold00001 | 138180 | 139326 | contig00045 | 45.8 |
| scaffold00001 | 139327 | 139493 | 167 | |
| scaffold00001 | 139494 | 143023 | contig00046 | 59.0 |
| scaffold00001 | 143024 | 143344 | 321 | |
| scaffold00001 | 143345 | 145519 | contig00047 | 50.5 |
| scaffold00001 | 145520 | 145963 | 444 | |
| scaffold00001 | 145964 | 147437 | contig00048 | 62.7 |
| scaffold00001 | 147878 | 149715 | contig00049 | 43.0 |
| scaffold00001 | 149716 | 150252 | 537 | |
| scaffold00001 | 150253 | 153703 | contig00050 | 46.0 |
| scaffold00001 | 153886 | 156279 | contig00051 | 57.8 |
| scaffold00001 | 156280 | 157134 | 855 | |
| scaffold00001 | 157135 | 158237 | contig00052 | 49.0 |
| scaffold00001 | 158238 | 158257 | 20 | |
| scaffold00001 | 158258 | 159199 | contig00053 | 57.1 |
| scaffold00001 | 159200 | 160103 | 904 | |
| scaffold00001 | 160104 | 162475 | contig00054 | 77.6 |
| scaffold00001 | 162476 | 162596 | 121 | |
| scaffold00001 | 162597 | 164104 | contig00055 | 49.1 |
| scaffold00001 | 164105 | 164911 | 807 | |
| scaffold00001 | 164912 | 169282 | contig00056 | 72.0 |
| scaffold00001 | 169735 | 176015 | contig00057 | 63.8 |
| scaffold00001 | 176122 | 176532 | contig00060 | 53.2 |
| scaffold00001 | 176533 | 176565 | contig00061 | 661.9 |
| scaffold00001 | 176566 | 179967 | contig00062 | 53.3 |
| scaffold00001 | 179968 | 179987 | 20 | |
| scaffold00001 | 179988 | 187338 | contig00063 | 67.4 |
| scaffold00001 | 187359 | 194047 | contig00064 | 61.3 |
| scaffold00001 | 194048 | 195103 | 1056 | |
| scaffold00001 | 195104 | 198561 | contig00065 | 66.8 |
| scaffold00001 | 198562 | 198581 | 20 | |
| scaffold00001 | 198582 | 201758 | contig00066 | 51.0 |
| scaffold00001 | 201821 | 206390 | contig00067 | 53.1 |
| scaffold00001 | 206391 | 206460 | contig00068 | 32.1 |
| scaffold00001 | 206461 | 211615 | contig00069 | 61.6 |
| scaffold00001 | 211829 | 218639 | contig00070 | 62.5 |
| scaffold00001 | 218808 | 222451 | contig00071 | 66.6 |
| scaffold00001 | 222687 | 226918 | contig00072 | 58.1 |
| scaffold00001 | 226943 | 227830 | contig00073 | 66.2 |
| scaffold00001 | 228604 | 229727 | contig00074 | 75.4 |
| scaffold00001 | 229728 | 230193 | 466 | |
| scaffold00001 | 230194 | 231663 | contig00075 | 68.2 |
| scaffold00001 | 231664 | 231937 | 274 | |
| scaffold00001 | 231938 | 236672 | contig00076 | 60.0 |
| scaffold00001 | 236673 | 237014 | 342 | |
| scaffold00001 | 237015 | 238896 | contig00077 | 55.3 |
| scaffold00001 | 238935 | 243532 | contig00079 | 73.2 |
| scaffold00001 | 243548 | 243710 | contig00081 | 81.4 |
| scaffold00001 | 243711 | 243825 | contig00082 | 32.2 |
| scaffold00001 | 243826 | 244114 | contig00083 | 46.3 |
| scaffold00001 | 244115 | 245025 | contig00084 | 49.1 |
| scaffold00001 | 245026 | 245045 | 20 | |
| scaffold00001 | 245046 | 246103 | contig00085 | 61.1 |
| scaffold00001 | 246104 | 246730 | 627 | |
| scaffold00001 | 246731 | 248175 | contig00086 | 55.2 |
| scaffold00001 | 248176 | 248628 | 453 | |
| scaffold00001 | 248629 | 250673 | contig00087 | 50.7 |
| scaffold00001 | 250674 | 251274 | 601 | |
| scaffold00001 | 251275 | 252063 | contig00088 | 54.0 |
| scaffold00001 | 252064 | 252778 | 715 | |
| scaffold00001 | 252779 | 254265 | contig00089 | 43.0 |
| scaffold00001 | 254266 | 255133 | 868 | |
| scaffold00001 | 255134 | 256947 | contig00090 | 58.9 |
| scaffold00001 | 256948 | 256967 | 20 | |
| scaffold00001 | 256968 | 259485 | contig00091 | 49.1 |
| scaffold00001 | 259758 | 262735 | contig00092 | 51.2 |
| scaffold00001 | 262736 | 263294 | 559 | |
| scaffold00001 | 263295 | 265486 | contig00093 | **88.6** |
| scaffold00001 | 265507 | 267077 | contig00094 | **50.7** |
| scaffold00001 | 267582 | 272617 | contig00095 | **65.4** |
| scaffold00071 | 1 | 2753 | Contig00278#1 | **162.3** |
| scaffold00003 | 1 | 1363 | contig00097 | 48.8 |
| scaffold00003 | 1653 | 2614 | contig00098 | 67.7 |
| scaffold00003 | 2920 | 3896 | contig00099 | 53.2 |
| scaffold00003 | 3933 | 9607 | contig00101 | 51.7 |
| scaffold00003 | 9679 | 10931 | contig00102 | 60.2 |
| scaffold00003 | 10932 | 11108 | 177 | |
| scaffold00003 | 11109 | 12073 | contig00103 | 39.1 |
| scaffold00003 | 12074 | 13052 | 979 | |
| scaffold00003 | 13053 | 14107 | contig00104 | 59.2 |
| scaffold00003 | 14108 | 14561 | 454 | |
| scaffold00003 | 14562 | 15411 | contig00105 | 46.4 |
| scaffold00003 | 15412 | 15643 | 232 | |
| scaffold00003 | 15644 | 16727 | contig00106 | 53.6 |
| scaffold00003 | 16728 | 17067 | 340 | |
| scaffold00003 | 17068 | 18682 | contig00107 | 45.1 |
| scaffold00003 | 18683 | 20247 | 1565 | |
| scaffold00003 | 20248 | 21435 | contig00108 | 56.4 |
| scaffold00003 | 21436 | 21455 | 20 | |
| scaffold00003 | 21456 | 24914 | contig00109 | 81.9 |
| scaffold00003 | 24915 | 25218 | 304 | |

Figura 31: **Esquema del "scaffold" cloroplástico de *Trebouxia* sp. TR9.** En la parte izquierda se muestra una tabla de las posiciones de los "contigs" en cada uno de los tres "scaffolds" originales (fondo blanco) y las distancias entre ellos en N's (fondo negro). Cuando diferentes "contigs" contiguos son del mismo color, denota que dichos "contigs" se han unido en uno único. En la parte central es un esquema de los "contigs" en barras de colores donde los cuadrados amarillos son las posiciones de los oligos diseñados. En la parte derecha son las lecturas mapeadas en cada "contig".

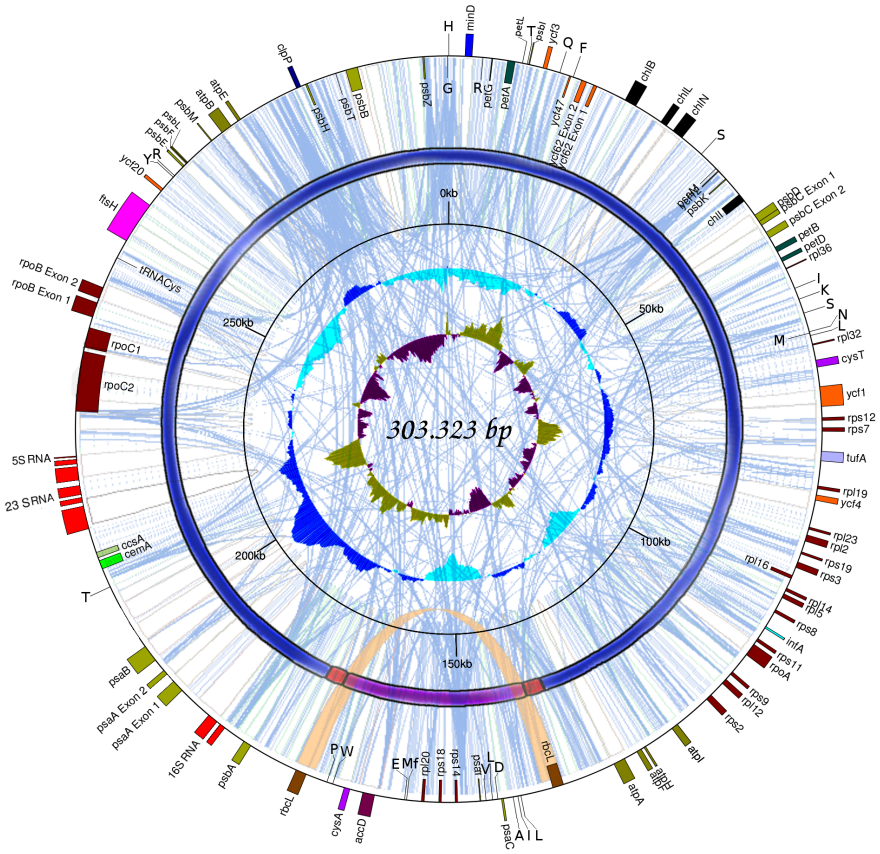



Figura 32: **Mapa del genoma cloroplástico de *Trebouxia* sp. TR9.** Los diferentes anillos, desde el interior hacia el exterior, muestran el sesgo en GC (GC skew), el contenido en GC, el tamaño genómico, y la posición de la LSC (Azul) la SSC (Violeta) y las IRs (Rojo). En el anillo más externo se representan los genes como cajas de colores, cuando se sitúan fuera y dentro del círculo se expresan en sentido horario y antihorario respectivamente. Lineas azules y naranjas en el fondo unen las repeticiones presentes en el cloroplasto.

teínas formando parte del proteosoma (gen *clpP*). Treinta y cuatro genes codifican proteínas que intervienen en el transporte electrónico fotosintético codificando componentes del fotosistema I (*psaA*, *B*, *C*, *I*, *M*) y del fotosistema II (genes *psbA*, *B*, *C*, *D*, *E*, *F*, *H*, *I*, *J*, *K*, *M*, *T*, *Z*, *ycf12*). Seis genes codifican componentes de la bomba de protones y síntesis de ATP (genes *atpA*, *B*, *E*, *F*, *H*, *I*). Cuatro genes codifican componentes que intervienen en la síntesis de clorofilas (genes *cemA*, *cysA*, *T*) y dos codifican proteínas relacionadas con la división celular (genes *ftsH*, *minD*). También se han identificado genes que codifican proteínas que participan en diferentes rutas del metabolismo como, por ejemplo, la fijación de $CO_2$ (gen *rbcL*) o/y la biosíntesis de ácidos grasos (*accD*). Por último, existen seis ORFs conservadas de funciones desconocidas por el momento (*ycf1*, *3*, *4*, *20*, *47*, *62*). Además, se han delimitado siete ORFs adicionales (*ORF321*, *ORF429*, *ORF451* se localizan dentro de intrones del gen que codifica el ARN 23S y codifican posibles "LAGLIDADG homing endocucleases" y los genes *ORF321*, *ORF429* y *ORF451* que no presentan similitud con ninguna proteína de la base de datos del NCBI). La mayor parte de los genomas de cloroplastos completamente secuenciados en Trebouxiophyceae presentan, al menos, un intrón (21 de 27) cuyo número máximo es de 28 en *Chlorsarcina brevispinosa* (N° de acceso KM462875), siendo *Trebouxia* sp. TR9 con un total de 12, una de las especies con mayor número de intrones. Todos ellos del tipo I y concentrados en una serie de genes como es el caso del gen que codifica la sub-unidad grande del ARN ribosomal, que alberga siete de ellos en su secuencia.

Cuadro 12: **Genes presentes en el genoma cloroplástico de *Trebouxia* sp. TR9**

| | |
|---|---|
| Transcription | |
| | *rpoA, B, rpoC1, C2* |
| Translation | |
| | *infA, tufA* |
| Small subunit | |
| | *rps2, 3, 7, 8, 9, 11, 12, 14, 18, 19* |
| Large subunit | |
| | *rpl2, 5, 12, 14, 16, 19, 20, 23, 32, 36* |
| | *clpP* |

| | |
|---|---|
| Chlorophyll biosynthesis | |
| | *chlB, I, L, N* |
| Photosystem I | |
| | *psaA, B, C, I, M* |
| Photosystem II | |
| | *psbA, B, C, D, E, F, H, I, K, L, M, T, Z, ycf12 (psb30)* |
| Cytochrome complex | |
| | *ccsA, petA, B, D, G, L,* |
| ATP synthase | |
| | *atpA, B, E, F, H, I* |
| Transport | |
| | *cemA, cysA, T* |
| Cell division | |
| | *ftsH, minD* |
| Carbon fixation | |
| | *rbcL* |
| Fatty acid biosynthesis | |
| | *accD* |
| Conserved ORFs | |
| | *ycf1, 20, 3, 4, 47, 62* |
| Ribosomal | |
| | *rrnL, rrnS* |

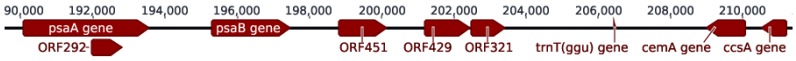



Figura 33: **Mapa genético de la región comprendida entre los genes *psaA* y *ccsA* del genoma de cloroplastos de *Trebouxia* sp. TR9.** Los genes y ORFs se muestran con flechas granate. Las posiciones en el genoma se indican en la parte superior.

| | |
|---|---|
| Transfer | *trnA-ugc, trnC-gca, trnD-guc, trnE-uuc, trnF-gaa, trnG-ucc, trnH-gug, trnI-cau, trnI-gau, trnK-uuu, trnL-caa, trnL-uaa, trnL-uag, trnMe-cau, trnMf-cau, trnN-guu, trnP-ugg, trnQ-uug, trnR-acg, trnR-tct, trnS-gct, trnS-uga, trnT-ggu, trnT-ugu, trnV-uac, trnW-cca, trnY-gua* |

Se han identificado una serie de pautas de lectura abierta (ORFs) que podrían codificar proteínas de más de 300 aminoácidos en una región del genoma de 8.947 nt comprendida entre las posiciones 197.493 y 206.440 entre los genes *psaB* y *trnT* (guu) (Figura 33). Estas ORFs denominadas *ORF41, ORF429* y *ORF321* que pueden codificar proteínas de 451, 429 y 321 aminoácidos, respectivamente. Búsquedas tBLASTn del posible CDS del *ORF451* han identificado una zona de 525 nt (175 aminoácidos) en el genoma de cloroplastos de *Dictyochloropsis reticulata* SAG215 (Nº acceso KM462860) localizada entre los genes *accD* y *psaA*, pero con unos niveles de similitud bajos (Cobertura 39 %, e-value 4e-07 e Identidad 29 %). Además, la traducción hipotética de esta región de *D. reticulata* contiene un codón de parada en la mitad de esta zona, indicando que esta región posiblemente no codifique ningún tipo de proteína. Las ORFs *ORF429* y *ORF321* no han mostrado ninguna similitud significativa con otro organismo depositado en las bases de datos del NCBI.

En lo relativo al uso de aminoácidos en las proteínas codificadas en el genoma cloroplástico de *Trebouxia* sp. TR9, los aminoácidos hidrofóbicos alifáticos Leucina (L) e Isoleucina (I), junto a los polares Serina (S) y Lisina (K) son los más abundantes. Por el contrario los aminoácidos hidrófobos Cisteína (C), Triptófano (W) y Metionina (M), junto al polar cargado positivamente Histidina (H), son los menos abundantes (Figura34).

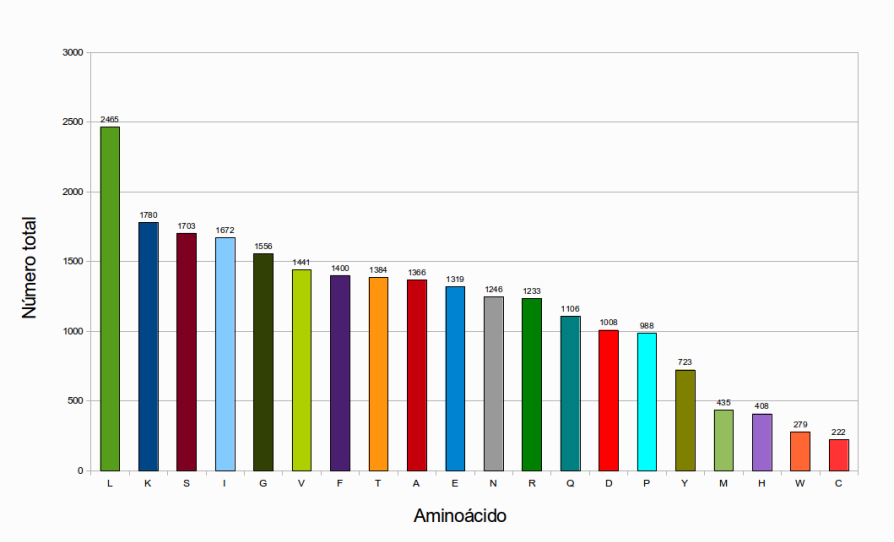



Figura 34: **Número total de cada aminoácido codificado en el genoma cloroplástico de *Trebouxia* sp. TR9.** Alanina (A), Cisteína (C), Ácido aspártico (D), Ácido glutámico (E), Fenilalanina (F), Glicina (G), Histidina (H), Isoleucina (I), Lisina (K), Leucina (L), Metionina (M), Asparagina (N), Prolina (P), Glutamina (Q), Arginina (R), Serina (S), Treonina (T), Valina (V), Triptófano (W), Tirosina (Y).

*Comparación del genoma cloroplástico de Trebouxia sp. TR9 con algas de la división Chlorophyta*

El tamaño del genoma cloroplástico de *Trebouxia* sp. TR9, en el contexto de las algas verdes de la división Chlorophyta, ocupa las primeras posiciones entre los publicados en el NCBI, situándose en tercer lugar después de *Prasiolopsis* sp. y *Floydiella terrestris* (Nº acceso NC_018569 y NC_014346, respectivamente). Al igual que en el caso de las algas de la clase Trebouxiophyceae, la longitud de la parte codificante es constante (unas 10 Kb) salvo en las algas pertenecientes a Mamiellophyceae, en las que se encuadran los eucariotas más pequeños (Figura 35). De modo que, en este caso, las diferencias de tamaño se deben, en mayor o menor medida, a la expansión de la parte no codificante. El genoma más compacto de las algas de Chlorophyta cuyas secuencias están disponibles en el NCBI, se corresponde con el del alga parásita *Helicosporidium* sp. (Nº acceso NC_008100) que ha perdido la mayoría de los genes relacionados con la maquinaria fotosintética.

En la tabla 13 se muestra un estudio comparativo de la presencia/ausencia de los genes que codifican proteínas en el genoma plastidial de *Trebouxia* sp. TR9 y especies de algas representativas de la división Chlorophyta. Los genes de función desconocida *ycf47* e *ycf62* solamente están presentes en muy pocas algas verdes. El gen *ycf47* se encuentra en las algas de la clase Trebouxiophyceae como *Trebouxia* sp. TR9, *Chlorella variabilis* (Nº acceso NC_015359), *Coccomyxa* sp. C-

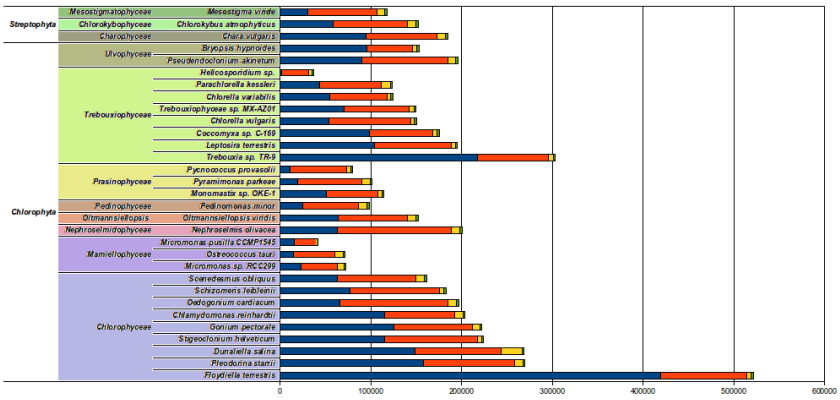
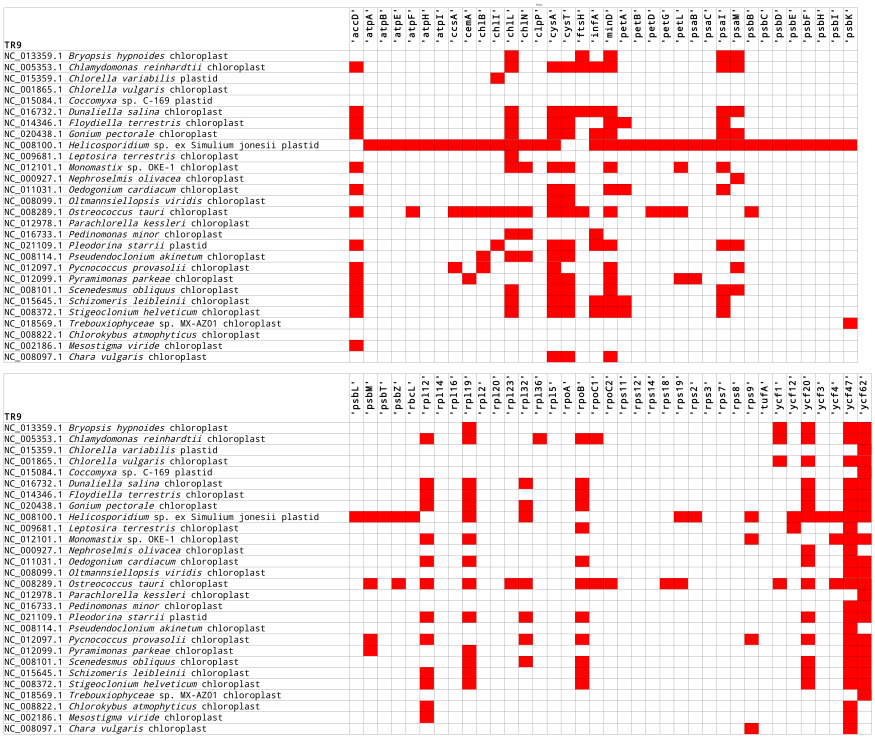



Figura 35: **Longitud en pares de bases de cloroplastos de algunas especies representativas de algas verdes de la división Chlorophyta.** En azul oscuro se indica la parte no codificante. En rojo, amarillo y verde se indican las proporciones codificantes de proteínas (CDSs), ARNs ribosomales (ARNrs) y ARNs de transferencia (ARNts), respectivamente.

160 (Nº acceso NC_015084), *Parachlorella kessleri* (Nº acceso FJ968741) y Trebouxiophyceae sp. MX-AZ01 (Nº acceso NC_018569), la clase Pedinophyceae como *Pedinomonas minor* y *P. tuberculata* (Nº acceso FJ968740 y NC_025530), la clase Prasinophyceae como *Nephroselmis astigmatica* (Nº acceso NC_024829) y la clase Ulvophyceae como *Bryopsis plumosa*, *Gloeotilopsis sterilis* y *Tydemania expeditionis* (Nº acceso NC_026795, NC_025538 y NC_026796). Sin embargo, este gen no se ha encontrado en ningún alga de la clase Chlorophyceae. El gen *ycf62* está presente en las algas de la clase Trebouxiophyceae como *Trebouxia* sp. TR9 y *Leptosira terrestris* (Nº acceso NC_009681), la clase Chlorophyceae como *Ettlia pseudoalveolaris* (Nº acceso NC_025532), la clase Pedinophyceae como *Pedinomonas tuberculata* (Nº acceso NC_025530), la clase Prasinophyceae como *Nephroselmis olivacea* y *Picocystis salinarum* (Nº acceso NC_024829 y NC_024828) y la clase Ulvophycea como *Pseudoclonium akinetum* y *Gloeotilopsis sterilis* (Nº acceso NC_008114 y NC_025538). Este gen también se encuentra en el genoma de cloroplastos de algas de la división Streptophyta como *Mesostigma viride* (Nº acceso NC_002186), *Chara vulgaris* (Nº acceso NC_008097) y *Chlorokibus atmophyticus* (Nº Acceso NC_008822). Así mismo, el gen *ycf20* también de función desconocida, se encuentra ausente en el genoma de cloroplastos de un elevado número de algas. Otros genes de funciones conocidas como los genes *cysA* y *cysT,* relacionados con el transporte transmembrana, así como *ftsH* y *minD,* relacionados con la división celular, no aparecen en el genoma de cloroplastos de gran parte de algas verdes. Algunas proteínas que forman parte de la subunidad grande de los ribosomas cloroplásticos (*rpl12*, *19* y *32*) no se encuentran en los genomas cloroplásticos de varios taxones de algas

verdes, así como el gen *rpoB*, que codifica una de las sub-unidades de la ARN polimerasa y genes de proteínas del fotosistema I como *psaI* y *psaM*. También el gen *chlL*, que está implicado en la síntesis de clorofila, está ausente en el genoma de cloroplastos de un elevado número de algas.




Cuadro 13: **Genes que codifican proteínas en el genoma cloroplástico de *Trebouxia* sp. TR9 y otras algas verdes.** Los genes de *Trebouxia* sp. TR9 presentes o ausentes en otros genomas cloroplásticos de algas verdes, se muestran en color blanco y rojo respectivamente.

*Comparación del genoma cloroplástico de Trebouxia sp. TR9 con el de otras algas de la clase Trebouxiophyceae (core Trebouxiophyceae)*

El tamaño del genoma cloroplástico de *Trebouxia* sp. TR9, es muy variable si comparamos los de 24 algas de la clase Trebouxiophyceae cuyas secuencias se encuentran actualmente disponibles en GenBank (Tabla 36). Los tamaños oscilan entre 306.152 nt en *Prasiolopis* sp. (Nº acceso KM462862) y 94.206 nt en *Choricystis parasitica* (Nº acceso KM462878), siendo el tamaño del genoma cloroplástico de *Trebouxia* sp. TR9 el segundo mayor (Figura 36). En la Figura 36, se puede observar que la longitud de la parte codificante es constante (unas 10Kb) de modo que las diferencias en los tamaños se deben, en mayor o menor medida, a la mayor o menor expansión de las regiones no codificantes.

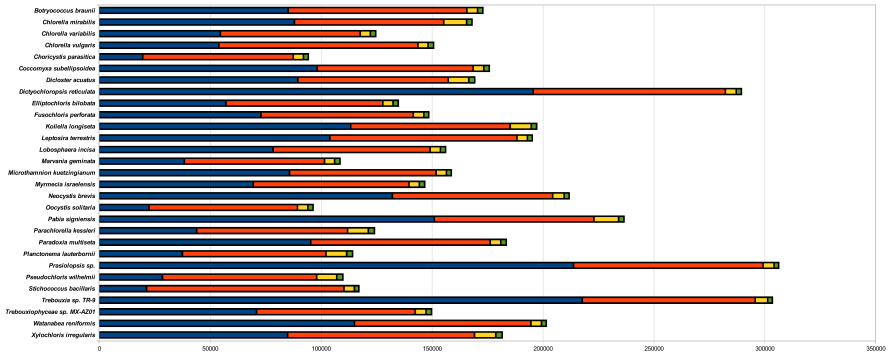



Figura 36: **Longitud en pares de bases del genoma de cloroplastos de algas verdes de la clase Trebouxiophyceae.** En azul oscuro se indica la parte no codificante, en rojo, amarillo y verde se muestran las porciones del genoma que codifican CDS's, ARNs y de ARNts, respectivamente.

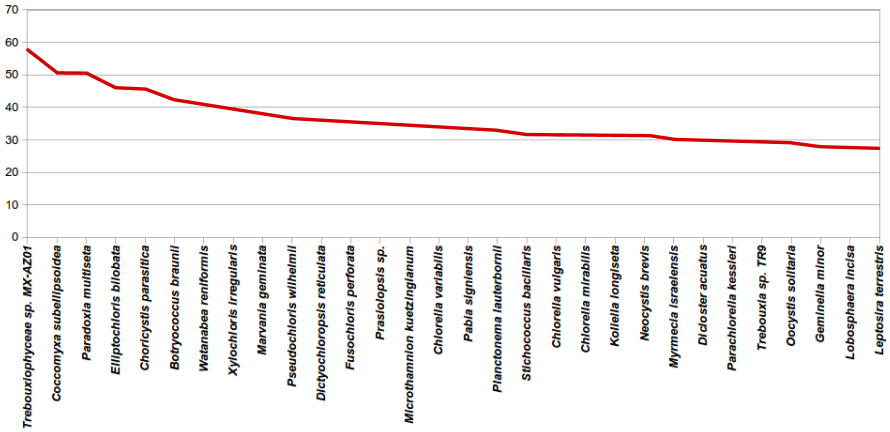



Figura 37: **Porcentaje de Guanina y Citosina de los genomas cloroplásticos de algas verdes de la clase Trebouxiophyceae.**

El contenido en GC es también muy variable (Figura 37) desde un 57,7 % en Trebouxiophyceae MXAZ01 (Nº acceso NC_018569) hasta un 27,3 % en *Leptosira terrestris* (Nº acceso EF506945), siendo en *Trebouxia* sp. TR9 de un 31,9 %, valor que se encuentra ligeramente por debajo de la media (36 %).

En la Tabla 13 se muestra un estudio comparativo de la presencia/ausencia de los genes que codifican proteínas en el genoma plastidial de *Trebouxia* sp. TR9 y otras algas del grupo conocido como "core Trebouxiophyceae" sensu Leliaert *et al.* (2012). Del total de 79 genes de la Tabla 13, solamente 15 están ausentes en no más de cinco especies de algas. Los genes *ycf47* e *ycf62* tan solo están ausentes en cinco especies de las 34 estudiadas mientras que el gen *ycf20* está presente en todas las algas estudiadas. Esta situación contrasta con la de las algas del resto de la división Chlorophyta en las que la presencia de estos genes en el genoma cloroplástico es muy poco frecuente. El

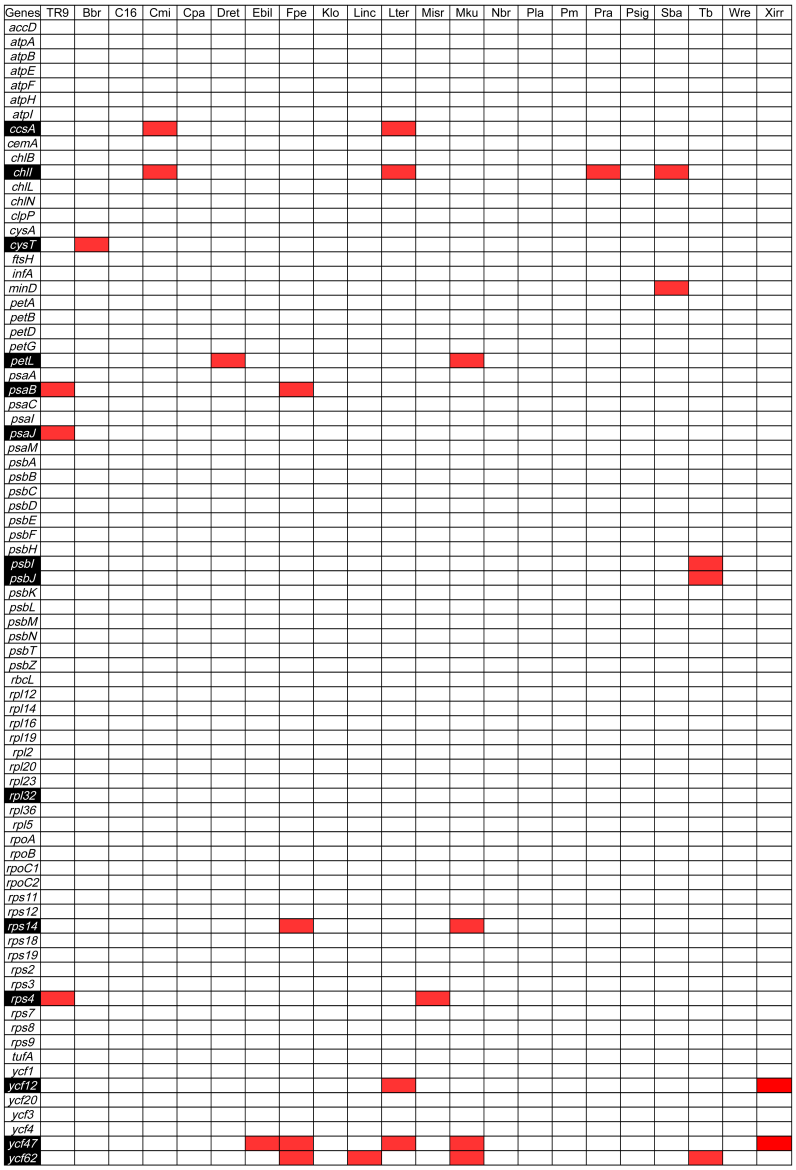
gen *chlI* relacionado con la síntesis de clorofila que está presente en casi todas las algas de Chlorophyta, está ausente en al menos cuatro especies de las Trebouxiophyceae. Los genes codificantes de proteínas ribosomales ausentes en los genomas de cloroplastos de algunas especies de Trebouxiophyceae son tanto de la sub-unidad pequeña (*rps4* y *14*) como de la grande (*rps32*). También faltan algunos genes que codifican proteínas de ambos fotosistemas (*psbA*, *psaJ*, *psbI* y *psbJ*).

La implementación del programa MAUVE entre los genomas de cloroplastos de tres algas pertenecientes a los géneros *Coccomyxa*, *Dictyochloropsis* y *Trebouxia* dio como resultado el gráfico de la Figura 38. Esta figura, representa las regiones homólogas que no presentan reordenamientos importantes. Entre los cpDNA de *Trebouxia* sp. TR9 y *Dictyochloropsis reticulata* existen 36 bloques según el cálculo DJC ("Double Cut and Join") con un valor de distancia DCJ de 33. Entre los cpDNAs de *Trebouxia* sp. TR9 y *Coccomyxa subellipsoidea* existen 48 bloques con un valor de distancia DCJ de 44. Entre los cpDNAs de *Coccomyxa subellipsoidea* y *Dictyochloropsis reticulata* existen 43 bloques con un valor de distancia DCJ de 42. Estos resultados no son consistentes con las relaciones filogenéticas entre los tres géneros (Lemieux *et al.* , 2014), como tampoco lo es el tamaño del genoma. Cabe señalar que tanto *Trebouxia* sp. TR9 como *Dictyochloropsis reticulata* son ficobiontes liquénicos mientras que *Coccomyxa subellipsoidea* es de vida libre. Por otra parte, estas diferencias en cuanto a la sintenia, no son sorprendentes dada la compleja arquitectura de los cpDNAs de algas verdes en comparación con los de plantas traqueofitas.

### *Identificación de intrones y Homing endonucleasas*

Se han localizado un total de doce intrones distribuidos entre los genes *rrnL*, *rrnS*, *psaA*, *psbC* y *rpoB*. Todos ellos son intrones del tipo I, siendo el gen que más intrones presenta el gen *rrnL*, que codifica el ARN 23S, con 7 intrones. El número de intrones del gen *rrnL* de *Trebouxia* sp. TR9 es solamente superado por el de *Floydiella terrestris* (Nº acceso NC_014346) con ocho intrones e igualado con siete intrones por *Dunaliella salina* (Nº acceso NC_016732). Los tamaños de los intrones de *Trebouxia* sp. TR9 oscilan entre 630 nt en el caso del intrón del gen *rpoB* hasta 1.361 nt en el caso del cuarto intrón del gen *rrnL*. El gen que codifica el ARN ribosomal 16S posee un solo intrón de 1.077 nt al igual que en otras muchas algas verdes. Los genes *psaA* y *rpoB* que codifican proteínas del fotosistema I y de la ARN polimerasa, respectivamente, contienen un intrón. La presencia de un intrón de 1.251 nt en el gen *psaA* se ha encontrado en otras algas de la clase Trebouxiophyceae como por ejemplo *Paradoxia multiseta* (Nº acceso KM462879), *Prasiolopsis* sp. SAG 8481 (Nº acceso KM462862) y *Stichococcus bacillaris* (Nº acceso KM462864). Algas de otras clases como, por ejemplo, las Chlorophyceae, también tienen al menos diez especies con un intrón en el gen *psaA* como en *Dunaliella salina* (Nº

Cuadro 14: **Genes que codifican proteínas en el genoma cloroplástico de *Trebouxia* sp. TR9.** Los genes de *Trebouxia* sp. TR9 presentes o ausentes en otros genomas cloroplásticos de algas del grupo "core Trebouxiophyceae" se muestran en color blanco y rojo respectivamente.



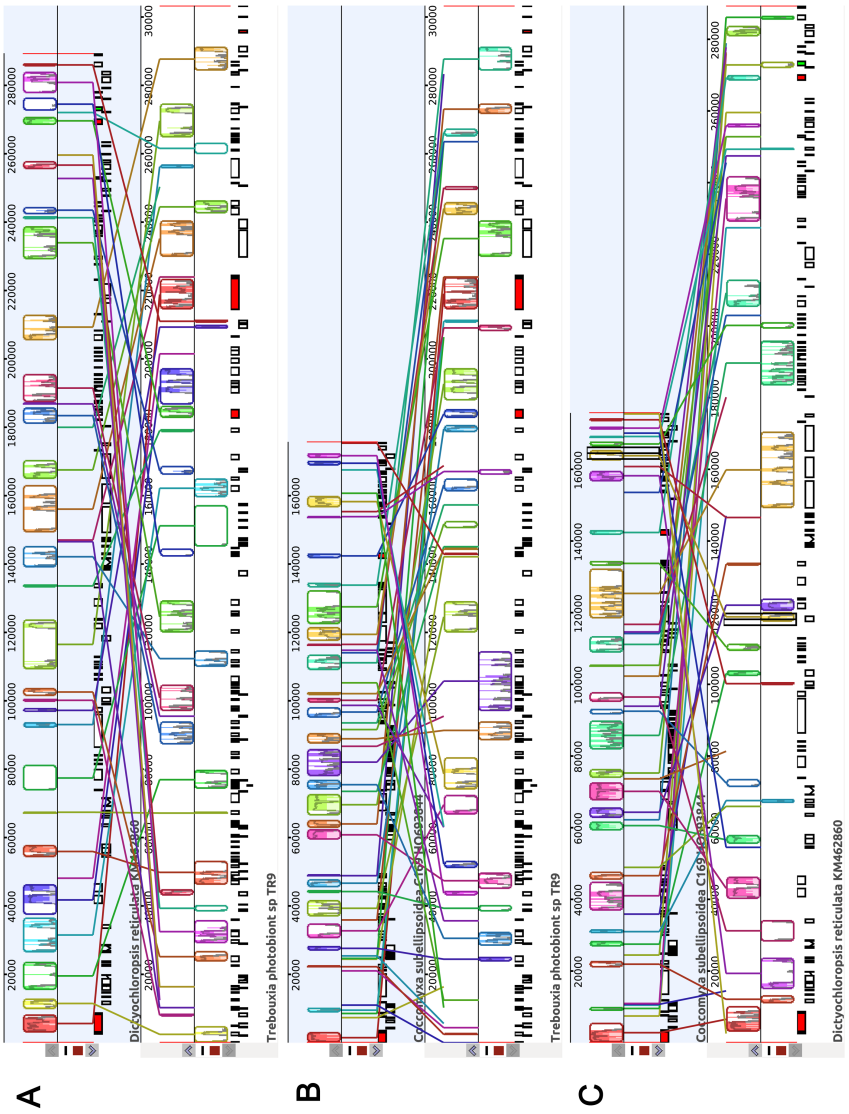



Figura 38: **Alineamientos de los genomas de cloroplastos de *Trebouxia* sp. TR9, *Dictyochloropsis reticulata* y *Coccomyxa subellipsoidea*.** El programa MAUVE (Darling *et al.* , 2004) se utilizó para alinear los genomas de cloroplastos de *Trebouxia* sp. TR9 y *Dictyochloropsis reticulata* (A), *Trebouxia* sp. TR9 y *Coccomyxa subellipsoidea* (B) y *Coccomyxa subellipsoidea* con *Dictyochloropsis reticulata* (C). Los bloques coloreados indican regiones con secuencias homólogas sin re-ordenamientos notables. Dentro de cada bloque se muestra el perfil de similitud de las secuencias. Las secuencias homólogas pero en sentido inverso, se muestran como bloques debajo de la línea central.

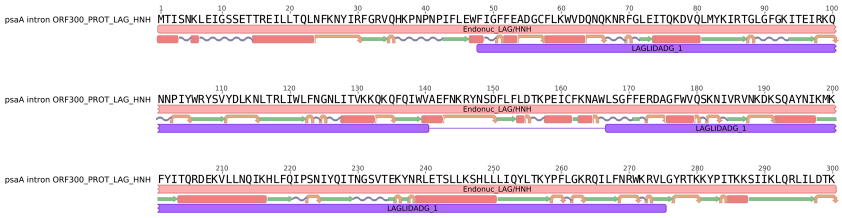



Figura 39: **Secuencia de aminoácidos y estructura secundaria de la proteína hipotética codificada en el intrón del gen *psaA*.** Las hélices alfa se muestran en rojo, las láminas beta en verde, los giros en amarillo y las espirales en azul. Debajo de la estructura secundaria se muestran los dominios reconocidos como propios de Homing endonucleasas de la familia1.

acceso NC_016732), *Pleodorina starrii* (Nº acceso NC_021109), *Schizomeris leibleinii* (Nº acceso NC_015645) y *Stigeoclonium helveticum* (Nº acceso NC_008372). El gen *psbC* presenta un intrón de 1.067 nt que también está, con cierta frecuencia, en otras especies de algas tanto de la clase Trebouxiophyceae como de las demás especies de la división Chlorophyta. En algunas especies de algas, el gen *psbC* posee incluso más de un intrón como, por ejemplo, *Pseudendochlonium akinetum* (Nº acceso NC_008114) con cuatro intrones y *Dunaliella salina* (Nº acceso NC_016732), *Oedogonium cardiacum* (Nº acceso NC_011031) y *Floydiella terrestris* (Nº acceso NC_014346) con dos intrones. Sin embargo, en los genomas completamente secuenciados de cloroplastos de otras especies de trebouxiofíceas, el gen *psbC* posee un solo intrón igual que en el caso de *Trebouxia* sp. TR9. En el gen *rpoB* es menos frecuente, hasta el momento, encontrar intrones y tan solo está en *Trebouxia* sp. TR9. El gen *ycf62* también posee un intrón que, al igual que el intrón del gen *rpoB*, se ha encontrado tan solo en *Trebouxia* sp. TR9. Como ya se ha mencionado anteriormente, este gen está presente en un gran número de especies de Trebouxiophyceae (Tabla 14) pero ausentes en muchas algas de otras clases de Chlorophyta (Tabla 14). Solamente algunos de los intrones del gen *rrnL* y el del gen *psaA* contienen una ORF que codifican una Homing endonucleasa de 300 aminoácidos con motivos LAGLIDADG y HNN (Figura 39).

### 4.3.2 *Discusión*

El genoma de cloroplastos de *Trebouxia* sp. TR9 fue secuenciado en el año 2012. El ensamblador utilizado para trabajar con las secuencias obtenidas de las plataformas ROCHE 454 GS FLX Titanium y ROCHE GS JUNIOR, no pudo ensamblar por completo el genoma, en parte por la presencia de una región invertida repetida que el ensamblador

colapsó en un único "contig" (con una cobertura de secuencias 4,6 veces mayor a la del resto de "contigs" de este genoma) y debido a que el genoma de cloroplastos de *Trebouxia* sp. TR9 contiene una alta fracción de secuencias repetitivas. Con la información de las secuencias "paired-end" y la ayuda guiada de PCRs, se pudieron conectar los 112 "contigs" del genoma en un "super-scaffold" que, finalmente, lo comprendían 50 "contigs". Este hecho parece ser común en este género de algas liquénicas, puesto que el genoma parcial de cloroplastos de *Trebouxia aggregata*, se encuentra en 40 "contigs" inconexos entre sí (Nº acceso EU123962-EU124002, Lemieux *et al.* 2014), pero que han sido útiles para realizar la filogenia de la clase Trebouxiophyceae (Lemieux *et al.* , 2014) y, también, han sido útiles para la anotación de ciertos genes del genoma de cloroplastos de *Trebouxia* sp. TR9. Con todo, el genoma de cloroplastos de *Trebouxia* sp. TR9 es el primero de éste género de algas verdes en ser secuenciado en toda su continuidad, aunque será necesario un mayor esfuerzo de secuenciación y/o análisis de las secuencias obtenidas en la plataforma Illumina para así, en el futuro, poder descifrar (si fuese posible) las zonas que en el "super-scaffold" obtenido están compuestas por "Ns" y por tanto obtener la totalidad de la secuencia de cloroplastos de *Trebouxia* sp. TR9.

*Tamaño del genoma, estructura y genes codificados*

En la actualidad se han secuenciado unos 60 genomas de cloroplastos de algas verdes de la división Chlorophyta. Las secuencias de estos genomas han mostrado una gran diversidad en cuanto al tamaño, arquitectura y composición de nucleótidos (Lang & Nedelcu, 2012; Leliaert *et al.* , 2012; Lemieux *et al.* , 2014). Muchos genomas de cloroplastos de algas verdes tienen una estructura característica en cuatro partes: dos regiones invertidas repetidas (IRa e IRb), una región de copia única larga (LSC) y otra de copia única corta (SSC). Si bien esta arquitectura parece ser bastante antigua, puesto que aparece en los linajes más basales de Chlorophyta, tales como la clase Pedinophyceae (p.e *Nephroselmis olivacea* (Turmel *et al.* , 1999a)), que es mantenida en la mayoría de plantas de la división Streptophyta. Sin embargo, el genoma de cloroplastos de muchas algas verdes carecen de este tipo de arquitectura (Wakasugi *et al.* , 1997; Bélanger *et al.* , 2006; De Cambiaire *et al.* , 2007; Turmel *et al.* , 2009; Brouard *et al.* , 2010).

Se ha obtenido la secuencia completa del genoma de cloroplastos de *Trebouxia* sp. TR9 que es de naturaleza circular y presenta un tamaño superior a 300 Kb. El tamaño de este genoma, en el contexto de las Chlorophyta, posee uno de los mayores tamaños, situándose en tercer lugar después de *Prasiolopsis* sp. y *Floydiella terrestris*. Como puede apreciarse en la Figura 32, tiene la estructura cuatripartita típica del genoma de cloroplastos de plantas terrestres. En el caso de

*Trebouxia* sp. TR9, las IRs incluyen un único gen, el *rbcL* que codifica la sub-unidad grande de la RuBisCO (Figura 32). En el conjunto del genoma, se han identificado un total de 108 genes de los que 77 codifican proteínas de funciones diversas, junto a tres posibles ORFs que podrían codificar proteínas mayores a 300 aminoácidos, pero que ninguna de ellas han podido identificarse con un grado alto de similitud con ninguna proteína/ORF de las depositadas en el NCBI. Esto pone de manifiesto la importancia de secuenciar un mayor número de genomas de cloroplastos de algas y otros organismos poco estudiados como musgos y helechos con el objeto de encontrar proteínas nuevas, de funciones por determinar, que podrían tener interés en biotecnología. Al igual que en el caso de las algas Trebouxiophyceae, la longitud de la parte codificante es constante (Unos 10 Kb) salvo en las algas pertenecientes a las Mamiellophyceae, donde se encuentran los eucariotas más pequeños (Figura 35) de modo que en este caso, las diferencias de tamaños se deben, en mayor o menor medida, a la expansión de la parte no codificante. El genoma más compacto de las Chlorophyta se corresponde con el del alga Trebouxiophyceae de vida parásita *Hellicosporidium* sp. (De Koning & Keeling, 2006) que ha perdido la mayoría de los genes relacionados con la maquinaria fotosintética. La comparación de los genomas de cloroplastos de tres especies de algas pertenecientes a los géneros *Trebouxia*, *Coccomyxa* y *Dictyochloropsis* que representan especies simbiontes de líquenes (Figura 38), indica la existencia de re-ordenamientos drásticos. Siendo mayores entre *Trebouxia* sp. TR9 y *Coccomyxa subellipsoidea* que entre *Trebouxia* sp. TR9 y *Dictyochloropsis reticulata.* Estudios comparados de regiones más locales entre taxones de trebouxiofíceas, indican que son mucho más comunes los reordenamientos en los genomas de cloroplastos de algas verdes si se comparan con los de plantas vasculares (Letsch & Lewis, 2012).

*Comparación del genoma cloroplástico de Trebouxia sp. TR9 con algas de la división Chlorophyta*

El estudio comparado de la presencia/ausencia de los genes que codifican proteínas en el genoma de cloroplastos de *Trebouxia* sp. TR9 y especies representativas de la división Chlorophyta (Tabla 13), indica que los genes de función desconocida *ycf47* e *ycf62* solamente están presentes en muy pocas algas verdes pertenecientes a esta división. El gen *yfc47* se encuentra en las algas de las clases Trebouxiophyceae, Pedinophycee, Prasinophyceae y Ulvophyceae. Sin embargo, no se ha encontrado en ningún alga de la clase Chlorophyceae. El gen *ycf62* está presente en algas de las clases Trebouxiophyceae, Chlorophyceae, Pedinophyceae, Prasinophyceae y Ulvophyceae. Este gen también se encuentra en el genoma de cloroplastos de algas de la división Streptophyta. Asimismo, el gen *ycf20*, de función desconocida, se encuentra ausente en el genoma de cloroplastos de un

elevado número de algas. Otros genes de funciones conocidas como los genes *cysA* y *cysT* relacionados con el transporte transmembrana así como los genes *ftsH* y minD relacionados con la división celular, no aparecen en el genoma de cloroplastos de gran parte de las algas verdes cuyas secuencias están disponibles en la base de datos del NCBI. Algunas proteínas que forman parte de la sub-unidad grande de los ribosomas de cloroplastos (*rpl12*, *19* y *32*), el gen *rpoB* que codifica una de las sub-unidades de la ARN polimerasa y genes de proteínas del fotosistema I como *psaI* y *psaM*, no se encuentran en los genomas plastidiales de varios taxones de algas verdes. También el gen *chlL* que está implicado en la síntesis de clorofila está ausente en el genoma de cloroplastos de un elevado número de algas. Nuestro estudio corrobora que el genoma de los cloroplastos de *Trebouxia* sp. TR9 conserva genes que en diferentes linajes de algas verdes han sido perdidos o que han pasado a ser parte de los genomas nucleares. Se puede postular la hipótesis de que en este alga, el control de la expresión de estos genes recae en el cloroplasto, posiblemente para poder actuar de manera más eficaz ante los rápidos cambios de hidratación/deshidratación que sufren estas microalgas en los talos de los líquenes, y así poder aprovechar rápidamente los escasos periodos óptimos para realizar los procesos de fotosíntesis y fijación de carbono.

*Identificación de introns y Homing endonucleasas*

La mayor parte de los genomas de cloroplastos de algas verdes completamente secuenciados presentan, al menos, un intrón cuyo número máximo es de 28 en *Chlorosarcina brevispinosa* (Nº acceso KM462875) siendo *Trebouxia* sp. TR9 una de las especies con mayor número de introns, con un total de 12, todos ellos de tipo I. Estos introns se distribuyen entre los genes *rrnL*, *rrnS*, *psaA*, *psbC* y *rpoB*, siendo el gen que más introns presenta el gen *rrnL* que codifica la sub-unidad 23S del ARN ribosomal con 7 introns. La presencia de introns en el gen *rpoB* es poco frecuente, encontrándose, hasta el momento, solamente en *Trebouxia* sp. TR9. Algunos de los introns presentes en el gen *rrnL* y *psaA* contienen ORFs que codifican "Homing endonucleasas" (HEs) de la familia LAGLIDADG (LHEs). El intrón presente en el gen *psaA* codifica una "Homing endonucleasa" de 300 aminoácidos con motivos LAGLIDADG y HNH (Figura 39). Algunos de los introns del gen *rrnL* codifican HEs de la familia LAGLIDADG. La mayoría de los introns de este gen fueron secuenciados en trabajos previos (del Campo *et al.* , 2009, 2010a) en *Trebouxia* sp. TR9 (Nº acceso EU600236), sin embargo, el intrón inserto en la posición 2.263 no se secuenció en dichos trabajos. Este intrón posee una LHE de 236 aminoácidos. Este intrón es similar al de *Trebouxia jamesii* UTEX 2233 (Nº acceso EU352794), inserto en la misma posición pero con delecciones que imposibilitan la unión de los cebadores cL2263F y cL2263

utilizados para la obtención del producto de amplificación por PCR que se obtiene con el ADN de *Trebouxia jamesii* UTEX2233 (Casano *et al.* , 2011). Esta discrepancia en la primera secuencia obtenida del gen *rrnL* de *Trebouxia* sp. TR9 puede explicarse por un error en el ensamblaje de las secuencias parciales obtenidas por el método Sanger o por la existencia de una variante de *Trebouxia* sp. TR9 sin el intrón. Seguramente, la explicación más probable es la primera, puesto que los cultivos de *Trebouxia* sp. TR9 utilizados en los trabajos anteriores al realizado en esta Tesis Doctoral (del Campo *et al.* , 2009, 2010a; Casano *et al.* , 2011; del Hoyo *et al.* , 2011; Álvarez *et al.* , 2012, 2014), utilizan clones del primer trabajo en el que se aisló esta alga del liquen *Ramalina farinacea* realizado por Gasulla *et al.* (2010). Además, la secuenciación ROCHE 454 realizada para la obtención del genoma de cloroplastos de *Trebouxia* sp. TR9, tiene una cobertura muchísimo más elevada de secuencias por nucleótido que la secuencia consenso (entorno a 60x, Figuras 30 y 31) obtenida por la secuenciación Sanger publicada en los trabajos de del Campo *et al.* (2009, 2010a).

*Relaciones filogenéticas entre las algas de la clase Trebouxiophyceae*

Las algas verdes representan un grupo muy diverso de organismos eucariotas fotosintéticos. Sin embargo, han sido mucho menos estudiadas que las plantas. Una de las principales razones de este desconocimiento general es la enorme dificultad del reconocimiento de taxones diferentes debido a los escasos caracteres morfológicos que las diferencian. La clase Trebouxiophyceae incluye organismos diversos con diferentes formas de vida tanto libre como en asociaciones con otros organismos que incluyen simbiosis de tipo mutualista y parásita. Las relaciones filogenéticas entre las algas de esta clase han sido recientemente establecidas en base a la secuencia de nucleótidos y aminoácidos derivadas de 79 genes cloroplásticos (Lemieux *et al.* , 2014). En la filogenia presentada por Lemieux *et al.* (2014) se distingue un clado integrado por 39 taxones que constituirían los que se llama "core trebouxiophyceans". Este grupo de algas son predominantemente terrestres. Dentro de este clado de "core trebouxiophyceans" el alga liquénica *Trebouxia agreggata* se agrupa con *Myrmecia israeliensis* en el orden *Trebouxiales* (Lemieux *et al.* , 2014). En líneas generales, nuestra filogenia basada en genes mitocondriales presentada en el apartado anterior, coincide con la de Lemieux *et al.* (2014) con la excepción de algunas relaciones entre las clases Prasinophyceae, Chlorophyceae, Trebouxiophyceae y Ulvophyceae. Probablemente, serán necesarios análisis adicionales con secuencias mitocondriales de más algas de esta división para resolver esta duda. Otra diferencia importante con respecto a lo obtenido por nosotros, se encuentra en la posición de la clase Pedinophyceae, que en la filogenia aquí presentada, aparece con valores de credibilidad superiores a los de Lemieux

*et al.* (2014), claramente relacionada con la clase Chlorophyceae y no con las Ulvophyceae.

## 4.4 GENOMA NUCLEAR

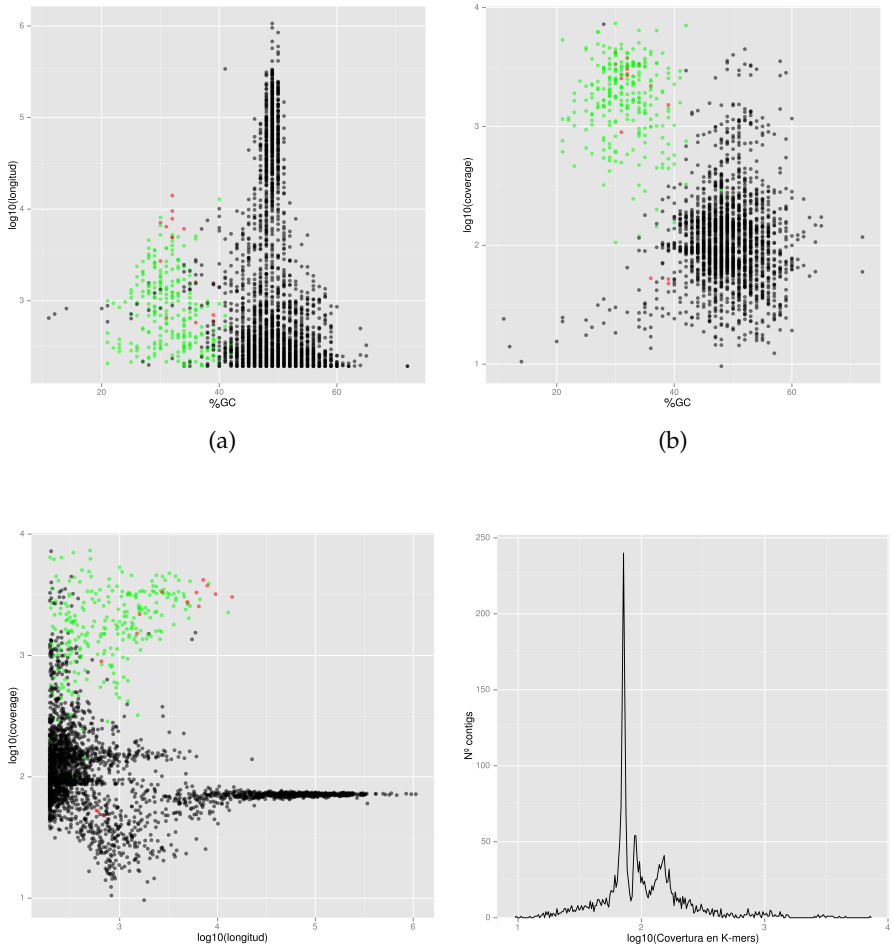
### 4.4.1 *Resultados*

*Estructura general del genoma*

En el transcurso de esta Tesis Doctoral se ha secuenciado parcialmente el genoma nuclear de *Trebouxia* sp. TR9 utilizando datos obtenidos por secuenciación masiva de ácidos nucleicos de las plataformas 454 (ROCHE 454 GS FLX Titanium y ROCHE 454 paired-end GS JUNIOR) e Illumina (Illumina Miseq paired-end). Un total de 57 ensamblajes fueron realizados con tamaños de k-mer impares desde el 21 al 133. El mejor ensamblaje obtenido estaba formado por 2.899 "contigs", de los cuales 16 y 259 pertenecían a los genomas mitocondrial y cloroplástico respectivamente. Los 2.626 "contigs" pertenecientes al genoma nuclear presentaron un N50 de 142.866 nt y un N95 de 21.727 nt. La media de longitud de todos los "contig" fue de 22.513,87 nt y el número total de residuos fue 59.121.427 nt, siendo el "contig" de mayor tamaño de 1.063.563 nt.

En la Figura 40 se muestran diferentes métricas del mejor ensamblaje obtenido. En la Figura 40 A se observa que, como en las imágenes internas de la Figura 12 correspondientes al contenido en Guanina y Citosina de las lecturas, aparecen dos picos diferenciados. El pico más alto (con "contigs" de mayor tamaño) corresponde a las lecturas nucleares y se encuentra alrededor del 50 %. El pico más bajo (con "contigs" de menor tamaño) corresponde a los genomas organulares cuyo %GC se encuentra entorno al 31 %. Al comparar el porcentaje de Guanina y Citosina frente a la cobertura de cada "contig" (Figura 40 C), de nuevo, se diferencian claramente los "contigs" provenientes de los orgánulos de los nucleares, tanto por su contenido diferente en Guanina y Citosina, como por la cobertura. La cobertura (en espacio de "K-mers") de los orgánulos es, en la mayoría de los "contigs" pertenecientes a la mitocondria y al cloroplasto, mayor a 1000x, mientras que en la mayoría de los "contigs" nucleares la cobertura es entrono al 70x (Figura 40 B, y C). En los casos donde la cobertura de los "contigs" nucleares es mayor, el tamaño de éstos es pequeño (Figura 40 C) por lo que podrían haber sido formados debido a la presencia de repeticiones o errores de secuenciación en las lecturas, que han llevado a un erróneo ensamblaje de éstas.

El programa RepeatMasker (Smit *et al.* , 1996) se ha utilizado para filtrar y anotar las secuencias de las repeticiones y secuencias de ADN de baja complejidad intercaladas en los "contigs" del ensamblaje nuclear de *Trebouxia* sp. TR9. En la Tabla 15 se presentan las repeticiones encontradas en el genoma de *Trebouxia* sp. TR9. Las repe-




(a) (b)

(c) (d)

Figura 40: **Métricas del mejor ensamblaje obtenido.** A) Longitud en base 10 de cada "contig" frente a su porcentaje de Guanina/Citosina. B) Porcentaje de Guanina/Citosina frente a la cobertura en base 10 de cada "contig". C) Cobertura de K-mers en base 10 frente a la longitud en base 10 de cada "contig". D) Frecuencia de "contigs" nucleares frente a su cobertura de K-mers en base 10. En los diagramas de puntos, los puntos de color verde, rojo y negro corresponden a "contigs" de procedencia cloroplástica, mitocondrial y nuclear respectivamente.

Cuadro 15: **Repeticiones obtenidas con RepeatMasker.**

| Clase | Tipo | Número de elementos | Longitud ocupada |
|---|---|---|---|
| **Retrotransposones sin LTR** | | | |
| | LINE1 | 14 | 1.375 |
| | LINE2 | 15 | 1.162 |
| | L3/CR1 | 11 | 822 |
| | R2/R4/NeSL | 13 | 3.759 |
| | RTE/Bov-B | 67 | 6.774 |
| | L1/CIN4 | 205 | 24.310 |
| **Retrotransposones con LTR** | | | |
| | ERVL | 21 | 3.641 |
| | ERVL-MaLRs | 1 | 52 |
| | ERV_classI | 2 | 87 |
| | Ty1/Copia | 119 | 25.489 |
| | Gypsy/DIRS1 | 339 | 69.582 |
| | Retroviral | 15 | 3.188 |
| **Transposones ADN** | | | |
| | Hobo-Activator | 1 | 50 |
| | Tc1- IS630-Pogo | 22 | 2.252 |
| | Otros (Mirage, P-element, Transib) | 23 | 1.922 |
| **Elementos de ADN** | | | |
| | TcMar - Tigger | 5 | 374 |
| **ARN no codificante** | | 34 | 8.752 |
| **Satélites** | | 4 | 488 |
| **Repeticiones Simples** | | 11.784 | 602.652 |
| **Baja complejidad** | | 648 | 35.624 |

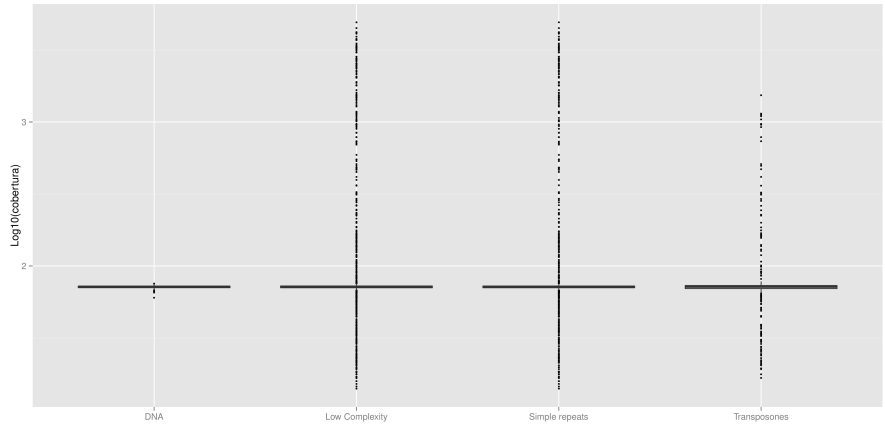



Figura 41: **Diagrama de cajas de las coberturas de los "contigs" que presentan diferentes tipos de repeticiones en el genoma nuclear de *Trebouxia* sp. TR9**

ticiones que ocupan el mayor número de nucleótidos dentro del genoma nuclear de *Trebouxia* sp. TR9 son las repeticiones simples, que ocupan un 1,01 % del total del ensamblaje. Dentro de los elementos móviles obtenidos, la mayoría de ellos pertenecen a retrotransposones con terminaciones terminales (LTR) con 497 elementos identificados y, en especial, los de tipo Gypsy/DIRS1 y Ty1/Copia (Tabla 15). La segunda clase más abundante de transposones son los de la clase de retrotransposones sin terminaciones largas con 325 elementos identificados diferentes. De los elementos repetitivos de ARN no codificante detectados por RepeatMasker, todas las identificaciones han sido las sub-unidades ribosomales nucleares. En los diagramas de cajas de la Figura 41 se puede observar que en cada tipo de repetición y transposón, la mayoría han sido detectados en "contigs" con un nivel de cobertura alrededor de 70x, aún así, los transposones, las repeticiones simples y de baja complejidad además han sido identificados en "contigs" de coberturas mucho mayores y menores. Estas secuencias han sido generadas por el ensamblador al colapsar cada una de estas repeticiones en un solo "contig" (Alta cobertura) o lecturas de estas repeticiones que el ensamblador ha colapsado en "contigs" de menor cobertura puesto que contienen errores de secuenciación.

Con la herramienta CEGMA (Parra *et al.* , 2009) se ha calculado qué porcentaje de un conjunto de 248 genes presentes en los genomas de todos los organismos eucariotas (CEGs) secuenciados se encuentran en el ensamblaje obtenido. Se han encontrado un 91 y 97 % del conjunto de 248 CEGs de forma completa y parcial respectivamente. Esta cifra indica que este ensamblaje comprende como mínimo, un 91 % del genoma nuclear de *Trebouxia* sp. TR9 y que podría ser del 97 % debido a la existencia de fragmentos pequeños mal ensamblados. Además, de los 236 "contigs" que obtuvieron uno o más genes del conjunto de CEGs, tan sólo cuatro con tamaños menores a 10.000

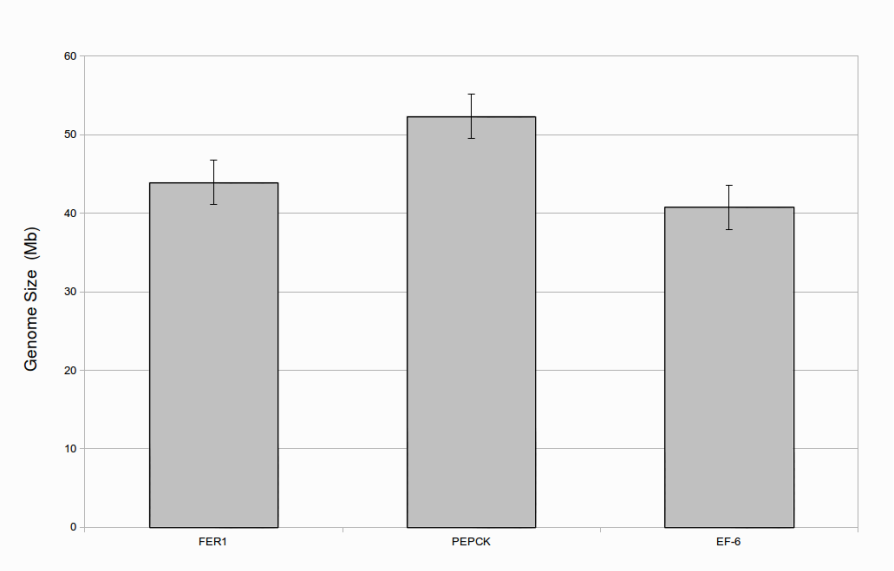



Figura 42: **Estimación del tamaño genómico de *Trebouxia* sp. TR9.** La altura de cada barra representa el tamaño calculado utilizando las parejas de oligos para los genes ferredoxin 1 (FER1), phosphoenol pyruvate carboxikinase (PEPCK) y elongation factor 6 (EF6).

nt (2.078 nt, 2.293 nt, 5.146 nt y 5208 nt ) presentaron alguno de estos genes. Por tanto, dentro del 24 % de los "contigs" restantes del ensamblaje (los mayores a 10.000 nt) se encuentra el 91 % del genoma nuclear de *Trebouxia* sp. TR9.

*Estimación del tamaño nuclear mediante Real-Time PCR*

El tamaño nuclear de *Trebouxia* sp. TR9 ha sido estimado con la metodología de Armaleo & May (2009) basada en las diferencias en el ciclo umbral (CT) de la reacción de PCR en Tiempo Real (RT-PCR) de diferentes concentraciones entre el amplificado del ADN genómico frente al amplificado por PCR convencional del mismo ADN genómico como molde con diferentes parejas de oligos. Los valores umbrales se utilizaron para calcular una curva estándar y su pendiente se utilizó para calcular la eficiencia de la RT-PCR. De todas las parejas de oligos utilizadas, sólo tres mostraron unas eficiencias adecuadas y/o no mostraban múltiples picos en las curvas de disociación (FER1, PPCK y EF6). De acuerdo con los datos de la Figura 42, la media del tamaño del genoma de *Trebouxia* sp. TR9 es entorno a las 40 - 50 mega bases.

Como complemento, se ha obtenido que el número de copias del gen del ARN ribosomal de *Trebouxia* sp. TR9 calculado con la fórmula de Wang *et al.* (2011) $2^{(\text{Ctref} - \text{Cttest})}$. Como puede observarse en la Figura 43, los ciclos umbrales para la pareja de oligos correspondientes al espaciador interno transcrito (ITS) del ARN ribosomal, son menores al ser un gen de copia múltiple. El gen FER1 se utilizó como gen

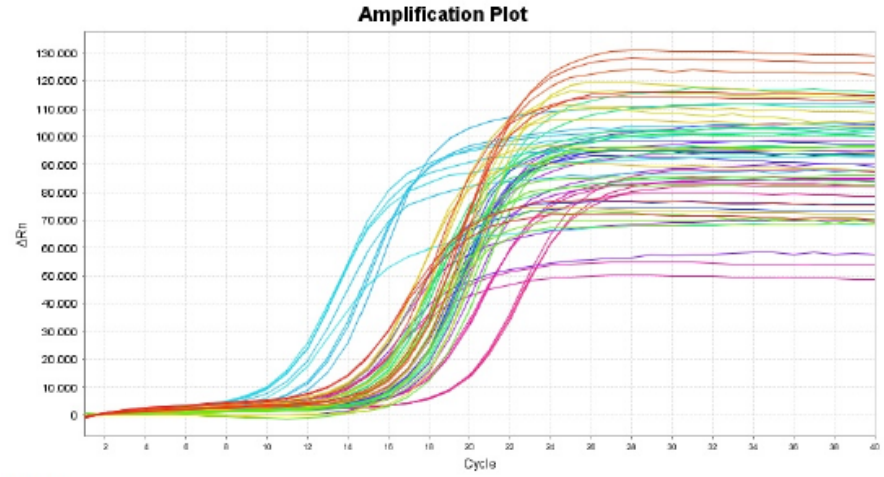



Figura 43: **Curvas de amplificación de diferentes parejas de oligos.** Las curvas azul celeste son las correspondientes a la pareja de oligos para el espaciador interno transcrito (ITS) del ARN ribosomal nuclear.

de referencia (Ctref) y el fragmento ITS1 como gen de ensayo (Cttest). El número de copias obtenido es cercano a 4 (3,91 ± 0,51).

*Anotación del genoma nuclear*

Una de las tareas más importantes dentro del contexto de la genómica junto al ensamblaje de ácidos nucleicos, es el proceso de anotación. Este proceso, en el caso de genomas nucleares de eucariotas, suele realizarse de forma automatizada ya que los genomas son de gran tamaño. La anotación se hace posible gracias a la estructura conservada de los genes que se encuentra mantenida a lo largo del árbol de la vida. Cuando no se poseen sistemas de predicción de genes basados en evidencias externas (como la obtención de las secuencias del ARN mensajero) es necesario recurrir a la predicción de genes "ab initio" junto a métodos comparativos de similitud con diferentes bases de datos biológicas. En el caso de eucariotas, la estructura exón - intrón hace de este proceso más complicado con métodos "ab initio". Para anotar el genoma nuclear de *Trebouxia* sp. TR9, se ha optado por la anotación "ab initio" con datos propios del genoma y su posterior comparación con diferentes bases de datos públicas.

En el proceso de predicción de genes "ab initio" del genoma nuclear de *Trebouxia* sp. TR9 se ha utilizado el programa AUGUSTUS (Stanke & Morgenstern, 2005). En primera instancia se recolectó un conjunto de genes para el entrenamiento de este software en relación a la estructura exón - intrón dentro de los genes de *Trebouxia* sp. TR9 y así obtener la señal de los sitios de empalme. Para ello, se utilizó la salida obtenida por CEGMA Parra *et al.* (2009) para encontrar las estructuras de los genes de proteínas "centrales" para

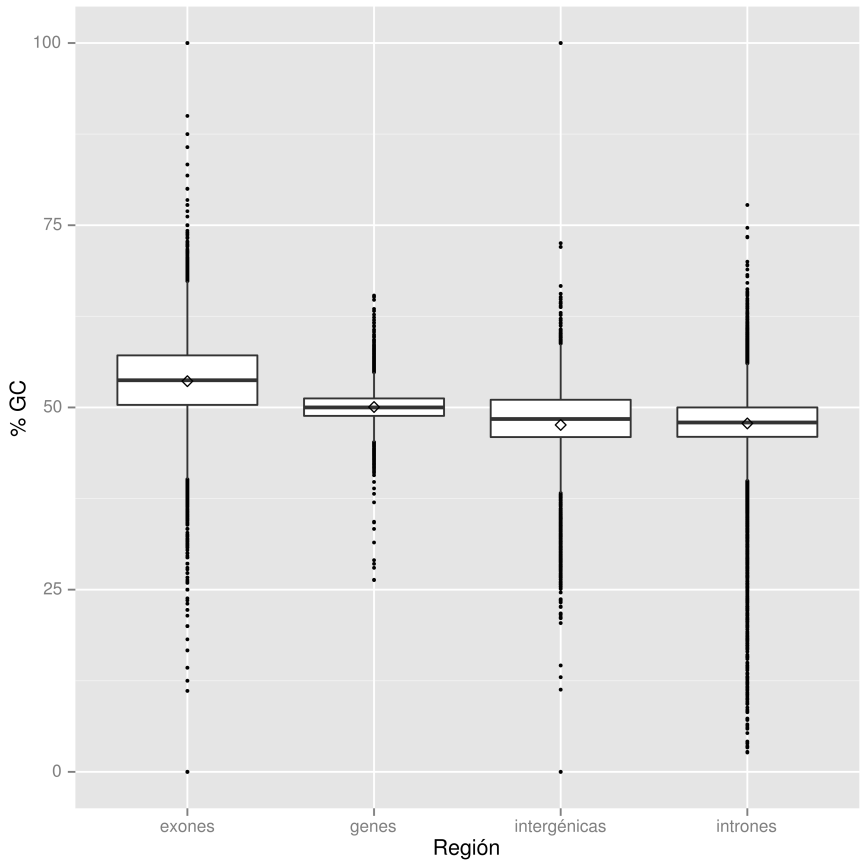



Figura 44: **Diagrama de cajas del contenido en Guanina y Citosina de los genes, intrones, exones y regiones intergénicas del genoma de *Trebouxia* sp. TR9.**

el metabolismo presentes en las secuencias genómicas de eucariotas. Se obtuvieron un total de 431 proteínas presentes en *Trebouxia* sp. TR9 con búsquedas realizadas con este software, mediante modelos probabilísticos complejos como son los modelos ocultos de Markov (HMMs). Con esta estrategia AUGUSTUS predijo 9.499 modelos de posibles genes que corresponden al 62,9 % del total del ensamblaje (37.430.807 de 59.502.099 nt totales). Éste número de predicciones de genes codificantes de proteínas es similar al de las algas trebouxiofíceas *Chlorella variabilis* NC64A (9.791) y *Coccomyxa subellipsoidea* (9.851) (Blanc *et al.* , 2010, 2012) y mayor al del alga simbionte del liquen *Cladonia grayii, Asterochloris* sp. (7.159) (`http://genomeportal.jgi-psf.org/Astpho1/Astpho1.info.html`). El tamaño del ensamblaje de los "contigs"/"scaffolds" mayores a 1.000 nt de *Trebouxia* sp. TR9 es de 58,8 Mb, siendo algo mayor al de las trebouxiofíceas *Chlorella variabilis* NC64A (46.2 Mb), *Coccomyxa subellipsoidea* (48,8 Mb), pero muy similar al de *Asterochloris* sp. (56.1 Mb).

La media de longitud, del número de exones y de intrones de los genes encontrados es 3.940,08 nt, 7,86 nt y 6,9 nt respectivamente. La

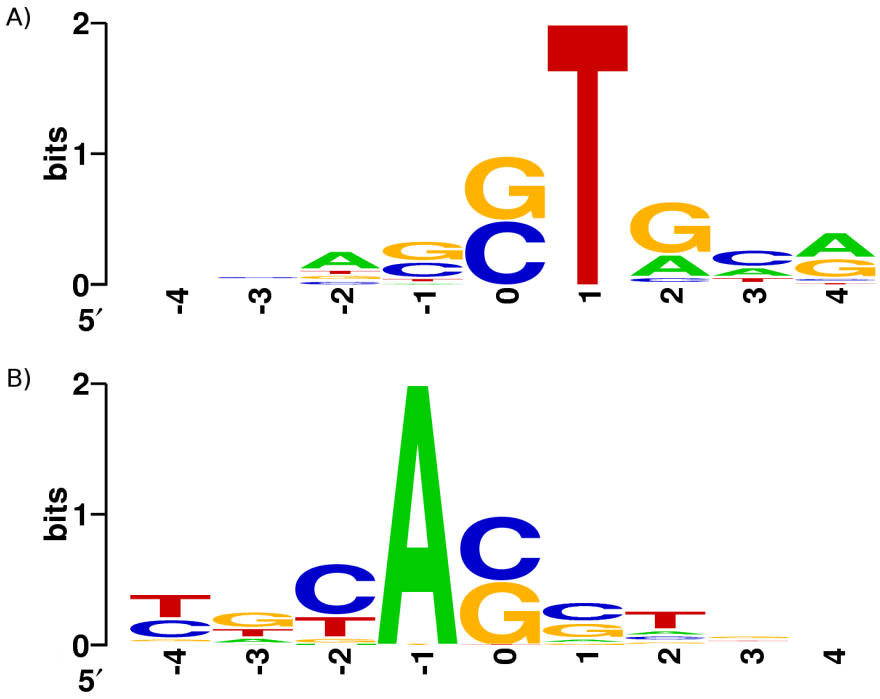



Figura 45: **Logotipos de las secuencias para los sitios de unión del espliceosoma de *Trebouxia* sp. TR9.** A) Sitios donantes de los intrones desde la posición -4 a la +4, siendo la posición o el inicio del intrón. B) Sitios aceptores de los intrones desde la posición -4 a +4, siendo la posición o el final del intrón.

densidad media de genes es de 5.42 genes por kilo base de genoma ensamblado. En la Figura 44 se observa que el tanto por ciento de Guanina y Citosina ( %GC) de los genes ,exones, intrones y en las regiones intergénicas tienen de media (± desviación estándar) 50,06 % (± 2,2), 53,605 (± 5,93), 47,80 % (± 3,99) y 47,60 % (± 7,6), respectivamente. Los logotipos de las secuencias de los sitios donde el espliceosoma elimina intrones de los pre-ARN mensajeros de los 65.295 intrones presentes en la anotación realizada por AUGUSTUS muestran que tanto el sitio donante de la parte 5′ como el aceptor de la parte 3′ son complementarios y que el nucleótido en la posición +1 del donante, junto al nucleótido en la posición -1 del aceptor (T/A, Figura 45) están muy conservados en los modelos génicos anotados, coincidiendo con los motivos de splicing de organismos eucariotas.

En lo relativo a los aminoácidos de los modelos de proteínas anotadas en el genoma nuclear de *Trebouxia* sp. TR9, los aminoácidos hidrofóbicos Alanina y Leucina junto a los hidrofílicos Serina y Glicina son los que se encuentran en mayor proporción mientras que los aminoácidos con menor proporción son el aminoácido hidrofóbico Triptófano y el hidrofílico Cisteína (Parte interna Figura 46). Este patrón es comparable al de diferentes algas verdes y plantas terrestres. En el caso del aminoácido hidrofílico Glutamina, *Trebouxia* sp.

TR9, presenta la mayor proporción de éste en los modelos proteicos obtenidos en comparación con el resto de plantas analizadas (Figura 46).

Tras la aproximación "ab initio" realizada y la obtención de las posibles proteínas codificadas en el genoma, se realizó un enfoque de genómica comparativa para poder mejorar la anotación y contrastarla con otras especies. Se realizaron búsquedas de similitud con el algoritmo BLAST (Altschul *et al.* , 1997) contra la base de datos de proteínas de referencia del NCBI (refseq_protein) y su posterior filtrado con la herramienta BLAST2GO (Conesa & Götz, 2008). Se encontraron 7.284 modelos proteicos anotados por AUGUSTUS con al menos un alineamientos significativo contra proteínas de la base de datos de proteínas de referencia del NCBI (BLASTp E-value = $1e^{-10}$ ), mientras que 2.215 no presentaron ningún alineamiento significativo ante este valor de corte. Las cuatro especies de algas verdes *Coccomyxa subellipsoidea*, *Chlorella variabilis*, *Volvox carteri* y *Chlamydomonas reinhardtii* se presentaron como mejor alineamiento en 3.046, 958, 526 y en 422 modelos proteicos respectivamente y las plantas del phylum Streptophyta *Physcomitrella patens* y *Selaginella mollendorfii* aparecieron como mejor alineamiento en 144 y 89 modelos proteicos de *Trebouxia* sp. TR9 respectivamente (Figura 47).

De los casi 9.500 modelos proteicos nucleares obtenidos en el proceso de anotación "ab initio", 6.364 fueron anotados con al menos un término de Gene Ontology (GO) y de ellos, 2.249 presentaron un número enzimático asociado. Tras utilizar la herramienta de anotación de dominios conservados del NCBI (CDD), 8.433 modelos presentaron al menos un dominio proteico conservado, 2.215 modelos proteicos que en el proceso de anotación de BLAST2GO no obtuvieron ningún tipo de anotación, 806 sí que presentaron algún motivo proteico conservado contra esta base de datos. De todos los modelos encontrados, aparecen 616 modelos proteicos con un solo exón, de los cuales, 370 no obtuvieron ningún alineamiento significativo en las búsquedas BLASTp, pero 85 obtuvieron un número GO y 2 un número enzimático. Además, 82 obtuvieron dominio conservado en CDD. Al carecer de información de tipo transcriptómico no es posible diferenciar si estos modelos son en realidad un gen sin intrones o sencillamente son anotaciones incorrectas.

La clasificación del enriquecimiento de términos GO presentes en la anotación de los modelos proteicos de *Trebouxia* sp. TR9 de acuerdo con su función distribuidas en procesos biológicos, funciones moleculares y en componentes celulares puede observarse en la Figura 48. Dentro de la categoría de función molecular, los términos GO "transferase activity" e "hydrolase activity" son los más representados. En la categoría de componentes celulares, los términos GO "protein complex", "membrane-enclosed lumen" y "nucleus" fueron los más representados. Finalmente, dentro de la categoría procesos bio-

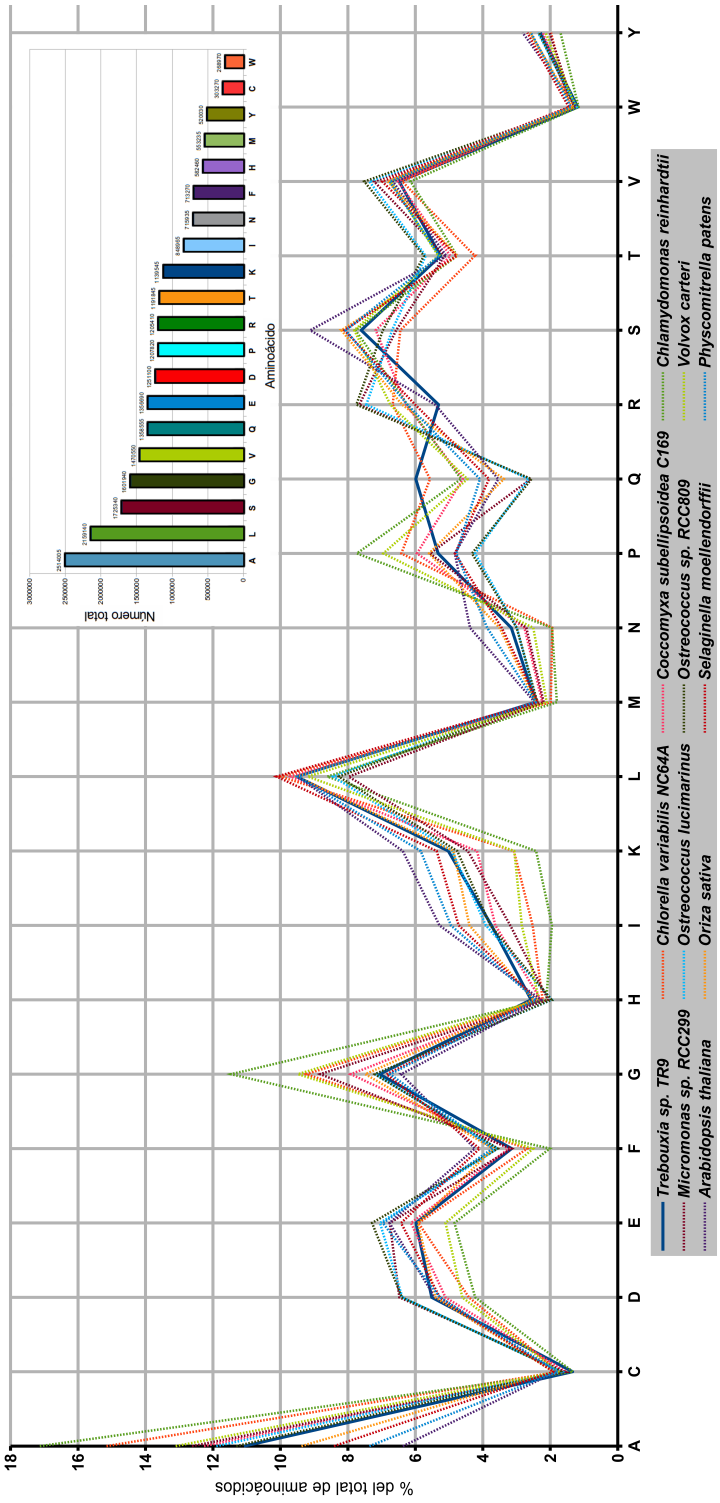



Figura 46: **Proporción de aminoácidos en los proteomas nucleares de *Trebouxia* sp. TR9 y diferentes algas verdes y plantas terrestres**. La figura interna muestra el número total de cada aminoácido codificado en el genoma mitocondrial de *Trebouxia* sp. TR9. Alanina (A), Cisteína (C), Ácido aspártico (D), Ácido glutámico (E), Fenilalanina (F), Glicina (G), Histidina (H), Isoleucina (I), Lisina (K), Leucina (L), Metionina (M), Asparagina (N), Prolina (P), Glutamina (Q), Arginina (R), Serina (S), Treonina (T), Valina (V), Triptófano (W), Tirosina (Y).

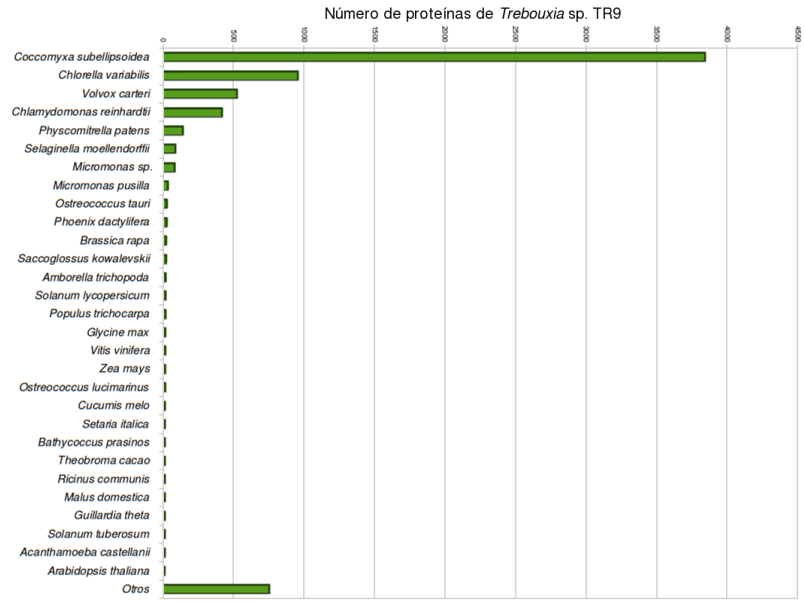



Figura 47: **Distribución de especies que dieron los mejores alineamientos contra los modelos proteicos de *Trebouxia* sp. TR9.**

lógicos, los términos GO que mayor enriquecimiento obtuvieron fueron el de "biosynthetic process", "nitrogen compound metabolic process", "small molecule metabolic process", "cellular macromolecule metabolic process", "gene expression", "catabolic process", "protein metabolic process" y "macromolecule modification".

*Dominios proteicos PFAM presentes en Trebouxia sp. TR9*

Se han obtenido un total de 10.922 dominios proteicos del PFAM de los cuales 3.156 resultaron únicos. Estos dominios estaban presentes en 6.544 modelos proteicos de *Trebouxia* sp. TR9. El porcentaje de modelos proteicos de *Trebouxia* sp. TR9 que presentaron al menos uno de estos dominios, es similar al de otros proteomas de diferentes organismos pertenecientes a las Viridiplantae. Las algas *Chlamydomonas reindhartii y Volvox carteri*, pertenecientes a la clase Chlorophyceae, junto al briófito *Physcomitrella patens*, son los que mayor proporción de proteínas sin modelos PFAM conocidos presentan en sus proteomas (Figura 49). Estas diferencias en el número de proteínas sin asignación de ningún motivo PFAM puede ser debida a que esta clase de algas verdes y el briófito han sufrido eventos evolutivos que han generado nuevos genes con motivos PFAM aún por descubrir.

La comparación de los motivos PFAM presentes en *Trebouxia* sp. TR9 con otras algas de la clase Trebouxiophyceae como *Asterochloris* sp., *Chlorella variabilis* y *Coccomyxa subellipsoidea* así como en el alga modelo *Chlamydomonas reindhartii* (Chlorophyceae), proporcionó los resultados que se indican a continuación (Figura 50). Del total de motivos PFAM presentes en los modelos proteicos de *Trebouxia* sp.

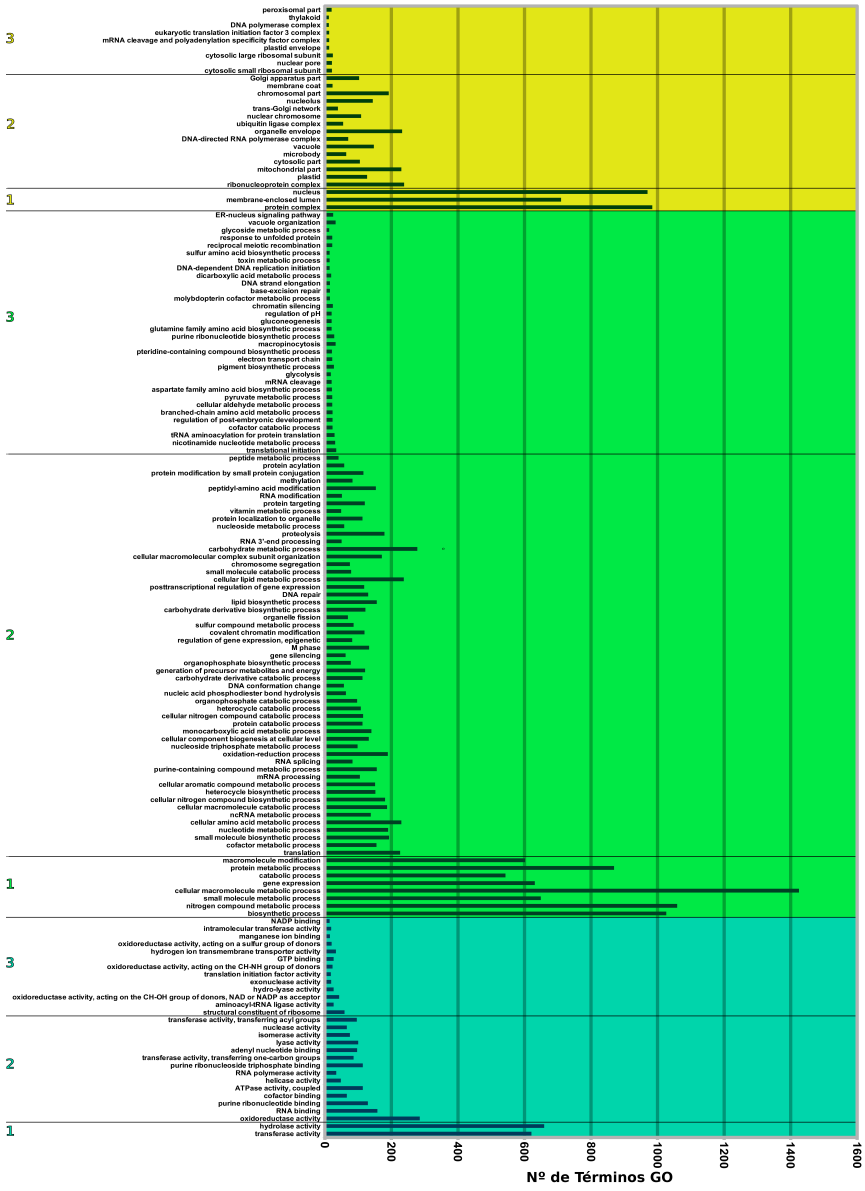



Figura 48: **Enriquecimiento de términos GO en la anotación de los modelos proteicos nucleares de *Trebouxia* sp. TR9.** En amarillo se muestran los componentes celulares, los procesos biológicos se marcan en verde y en azul turquesa las funciones moleculares. Los números 1, 2, y 3 a la izquierda de la figura indican funciones muy generales, generales y específicas de cada una de las tres categorías respectivamente.

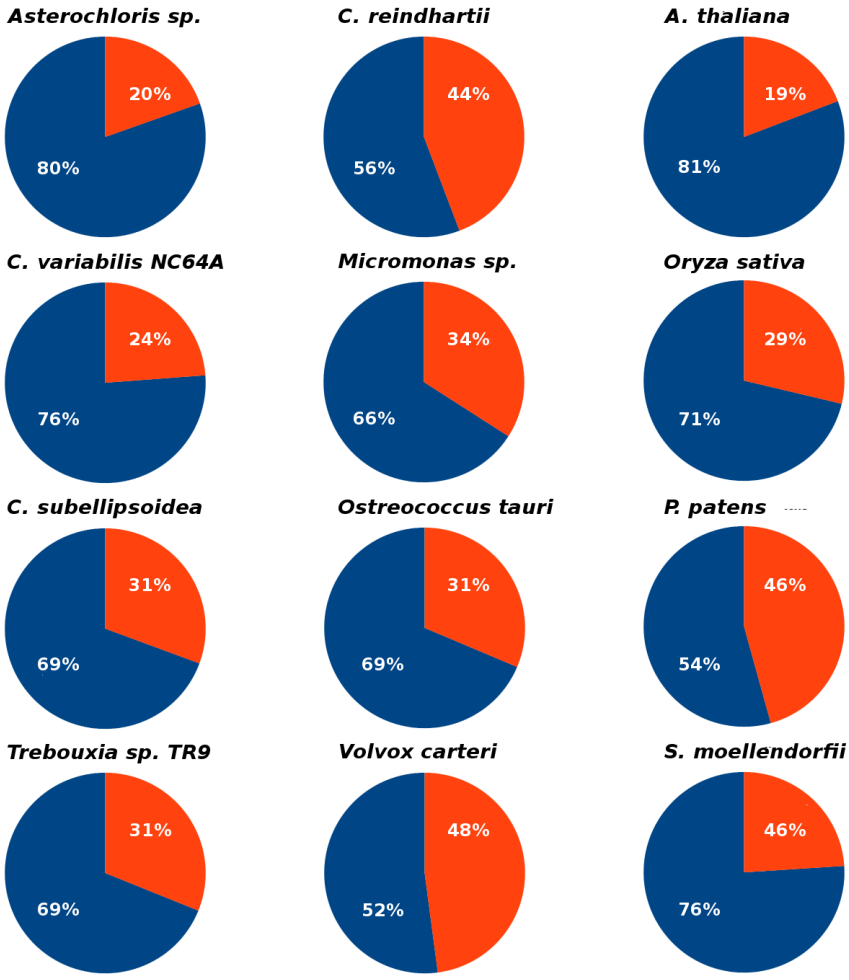



Figura 49: **Contenido de dominios PFAM en plantas.** Los colores azul y rojo son los porcentajes de proteínas que presentan o no dominios PFAM, respectivamente.

TR9, 2.147 son compartidos por todas las especies de algas estudiadas y 2.979 eran compartidos con al menos una de las especies de las algas estudiadas. Además, aunque no se han podido representar adecuadamente en la Figura 50, han aparecido 39 motivos PFAM que son compartidos entre *Trebouxia* sp. TR9, *Asterochloris* y *Coccomyxa*; 76 que son compartidos por *Trebouxia* sp. TR9, *Chlorella* y *Chlamydomonas* y, finalmente, 95 motivos que sólo están presentes en los modelos proteicos de *Trebouxia* sp. TR9.

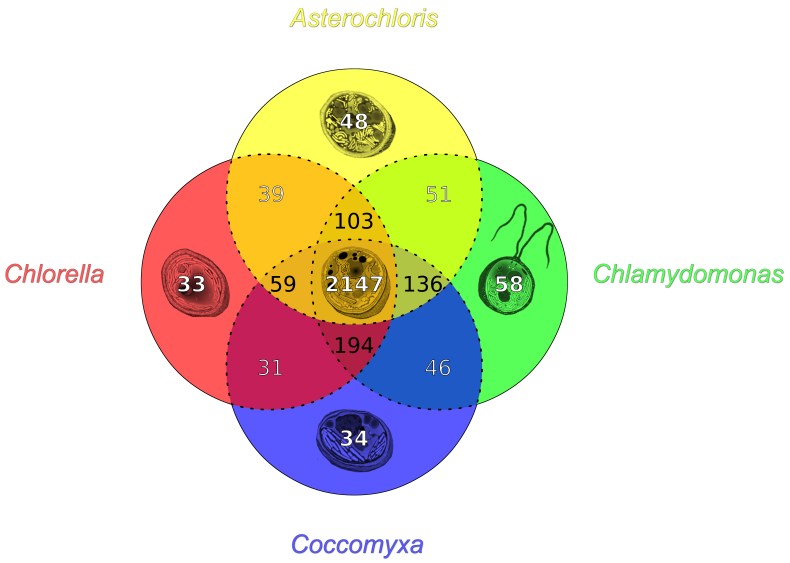



Figura 50: **Diagrama de Venn de los dominios PFAM compartidos entre *Trebouxia* sp. TR9, las algas *Asterochloris* sp., *Chlorella variabilis* NC64A, *Coccomyxa subellipsoidea* y *Chlamydomonas reindhartii*.**

Cuando, en las comparaciones se añadieron otros organismos además de los anteriores, como otras clases de algas verdes, el musgo *Physcomitrella patens*, el helecho *Sellaginella moellendorffii*, la monocotiledónea *Oryza sativa* y la dicotiledónea *Arabidopsis thaliana*, se detectaron motivos específicos de la división Chlorophyta (Tabla 16). Del total de 5.164 motivos únicos presentes en todas estas especies, 19 PFAM sólo estaban presentes en las algas verdes y de ellos, 4 eran dominios de función desconocida (DUF).

Por otra parte, en las cuatro plantas terrestres analizadas se encontraron 236 dominios PFAM, de los cuales 70 eran de función desconocida (DUF) y que no estaban presentes en ninguna de las algas analizadas. Muchos de estos dominios son factores de transcripción o de respuesta frente al estrés, factores hormonales (auxinas y citoquininas), enzimas del metabolismo de azúcares o modificadores de cromatina y del ciclo celular.

El análisis de los motivos PFAM propios de las algas liquénicas *Trebouxia* sp. TR9 y *Asterochloris* sp. que no están presentes en las algas

Cuadro 16: **Dominios proteicos PFAM específicos de Clorofitas.** TR9, *Trebouxia* sp. TR9; Asp, *Asterochloris* sp.; Csu, *Coccomyxa subellipsoidea*; Cva, *Chlorella variabilis*; Cre, *Chlamydomonas reindhartii*; Vca, *Volvox carteri*; Mcc, *Micromonas pusilla* CCMP1545; Mrc, *Micromonas pusilla* RCC299; Olu, *Ostreococcus lucimarinus*; Osp, *Ostreococcus* sp. RCC809; Ota, *Ostreococcus tauri*.

| Dominio PFAM | TR9 | Asp | Csu | Cva | Cre | Vca | Mcc | Mrc | Olu | Osp | Ota | Descripción |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PF00710 | 2 | 1 | 1 | 2 | 5 | 2 | 1 | 1 | 2 | 1 | 1 | Asparaginase |
| PF01082 | 1 | 1 | 1 | 1 | 5 | 2 | 1 | 2 | 1 | 2 | 2 | Copper type II ascorbate-dependent monooxygenase, N-terminal domain |
| PF01130 | 1 | 1 | 2 | 4 | 2 | 2 | 2 | 1 | 2 | 1 | 2 | CD36 family |
| PF01226 | 4 | 2 | 2 | 3 | 6 | 5 | 1 | 2 | 1 | 1 | 1 | Formate/nitrite transporter |
| PF01951 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | Archease protein family (MTH1598/TM1083) |
| PF02436 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | Conserved carboxylase domain |
| PF02656 | 2 | 3 | 3 | 4 | 5 | 5 | 2 | 2 | 1 | 1 | 1 | Domain of unknown function (DUF202) |
| PF03712 | 1 | 1 | 1 | 1 | 6 | 2 | 1 | 2 | 1 | 2 | 2 | Copper type II ascorbate-dependent monooxygenase, C-terminal domain |
| PF05517 | 5 | 4 | 3 | 5 | 11 | 6 | 8 | 5 | 2 | 2 | 2 | p25-alpha |
| PF05889 | 1 | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | Soluble liver antigen/liver pancreas antigen (SLA/LP autoantigen) |
| PF09418 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | Protein of unknown function (DUF2009) |
| PF11051 | 7 | 1 | 4 | 2 | 4 | 2 | 1 | 1 | 1 | 1 | 2 | Mannosyltransferase putative |
| PF11378 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | Protein of unknown function (DUF3181) |
| PF11397 | 1 | 4 | 5 | 2 | 4 | 2 | 2 | 3 | 1 | 1 | 1 | Glycosyltransferase (GlcNAc) |
| PF12345 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | Protein of unknown function (DUF3641) |
| PF13000 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | Acetyl-coenzyme A transporter 1 |
| PF13183 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 4Fe-4S dicluster domain |
| PF13472 | 9 | 3 | 12 | 29 | 17 | 12 | 4 | 3 | 4 | 2 | 2 | GDSL-like Lipase/Acylhydrolase family |
| PF14186 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | Cytoskeletal adhesion |

verdes ni en las plantas terrestres analizadas, *Physcomitrella patens*, *Oryza sativa* y *Arabidopsis thaliana* analizadas, puso de manifiesto la presencia de seis motivos PFAM exclusivos de estas algas. Las funciones que pueden atribuirse a las proteínas portadoras de estos motivos estaban relacionadas con el ciclo celular: proteínas de unión a ADN ("BRO family, N-terminal domain", PF02498 ), función kinasa relacionada con la apoptosis ("FAST kinase-like protein, subdomain 1", PF06743 ) y con canales de regulación mediada por calcio ("Polycystin cation channel", PF08016). También se han encontrado dos dominios presentes en las dos sub-unidades de la "Nitrile hydratase", encargada de la conversión de compuestos que contienen grupos nitrilo en amoniaco y ácidos orgánicos ("Nitrile hydratase, alpha chain", PF02979 y "Nitrile hydratase beta subunit", PF02211). Por último, se ha encontrado un dominio PFAM relacionado con las cápsidas de virus ("Large eukaryotic DNA virus major capsid protein", PF04451). Un análisis detallado de estos modelos proteicos mostraron que ambas proteínas contienen diversos intrones en sus secuencias.

Por último, se encontraron 25 motivos PFAM propios de *Trebouxia* sp. TR9, de los que 11 eran dominios de función desconocida (DUF), 4 relacionados con uniones a factores de transcripción ("NIF3 (NGG1p interacting factor 3)", PF01784; "Helix-turn-helix", PF08222; "S1-like", PF14444 y "Primase C terminal 2 (PriCT-2)", PF08707), 2 estaban relacionados con actividades kinasas y fosfatasas ("PrkA AAA domain", PF08298 y "Protein phosphatase 1 regulatory subunit 35 C-terminus" PF15503), 2 que podrían estar relacionados con la pared celular ("Fi-

bronectin type II domain", PF00040 y el dominio "Chitin synthase", PF03142) y 4 relacionados con virus ("KilA-N domain", PF04383; "Bacteriophage replication gene A protein", "Bacteriophage replication gene A protein", PF05840; "Microvirus H protein", PF04687 y "Phage Tail Collar Domain", PF07484). Al comprobar que había especies de gramíneas y fabáceas en las que también se encuentran los 2 últimos dominios PFAM propios de *Trebouxia* sp. TR9 ("RNA polymerase beta subunit external 1 domain", PF10385 y "Alpha helical coiled-coil rod protein (HCR)", PF07111), estos resultados se interpretaron como falsos positivos.

### 4.4.2 *Discusión*

Los genomas nucleares de las algas verdes están poco estudiados puesto que hasta mayo de 2015, tan sólo se habían publicado 8, de los cuales dos de clorofíceas: *Chlamydomonas reindhartii* (Merchant *et al.* , 2007) y *Volvox carteri* (Prochnik *et al.* , 2010); 4 de mamielofíceas: *Ostreococcus tauri* (Derelle *et al.* , 2006), *Ostreococcus lucimarinus* (Palenik *et al.* , 2007) y las cepas de *Micromonas pusilla* CCMP1545 y RCC299 (Worden *et al.* , 2009); 2 trebouxiofíceas: *Chlorella variabilis* (Blanc *et al.* , 2010) y *Coccomyxa subellipsoidea* ((Blanc *et al.* , 2012)). Además, junto a estas especies, el alga liquénica *Asterochloris* sp. tiene sus genomas secuenciados, pero no publicados (`http://genomeportal.jgi-psf.org/Astpho1/Astpho1.info.html`). Los datos genómicos y sus resultantes modelos génicos son fundamentales para el estudio de diferentes procesos evolutivos como, por ejemplo, la sintenia de genes entre especies, la información de señalización proteica (residente en la porción N-Terminal del péptido), la caracterización de las diferentes familias proteicas, el mapeo de transcritos de cADN o de péptidos obtenidos por espectrometría de masas, la predicción de análisis funcionales o para poder esclarecer los procesos de diversificación del árbol de la vida, junto a otras muchas aplicaciones biotecnológicas (Kim *et al.* , 2014; Bhattacharya *et al.* , 2015). El objetivo principal del trabajo realizado para la presente Tesis ha sido el de secuenciar y ensamblar el genoma nuclear del alga liquénica *Trebouxia* sp. TR9, obtener un primer borrador de éste y caracterizar los posibles modelos génicos presentes en él.

#### *Estructura general del genoma*

El borrador del genoma nuclear de *Trebouxia* sp. TR9 ha sido generado con secuencias obtenidas con las tecnologías de secuenciación ROCHE 454 GS FLX Titanium, ROCHE 454 GS JUNIOR "paired-end" e Illumina Miseq "paired-end" ensambladas con el ensamblador Velvet. De todos los ensamblajes realizados, el obtenido con un tamaño de 97 nucleótidos de K-mer fue el que mejores estadísticos tenía. Tamaños bajos de este factor (menores a 81) dieron como resultado

ensamblajes muy fragmentados que puede ser consecuencia de la presencia de repeticiones presentes en el genoma menores a este valor. La gran mayoría de "contigs"/"scaffolds" nucleares (menores a 10 Kb) presentaron un nivel de cobertura genómica muy alta que el ensamblador ha formado debido a regiones repetitivas del genoma, en cambio, en los "scaffolds" de mayor tamaño, su nivel de cobertura genómica se situó alrededor del 70x y su proporción de Guanina y Citosina fue de alrededor del 50 %, en concordancia con el de las lecturas originales. El ensamblaje del genoma nuclear de *Trebouxia* sp. TR9 contiene un número muy bajo de elementos móviles, repeticiones simples y zonas de baja complejidad (0,24 %, 1,01 % y 0,06 %, respectivamente) en comparación con la proporción de este tipo de elementos en otras algas verdes de la clase Trebouxiophyceae (12 % en el genoma de *Chlorella variabilis* NC64A y 7.2 % en *Coccomyxa subellipsoidea* C-169) o de plantas terrestres como *Arabidopsis thaliana* (del 20 al 30 % de su genoma) o *Triticum aestivum* (mayor al 90 % de su genoma). Este resultado indica que el contenido en secuencias repetidas de *Trebouxia* sp. TR9 probablemente se encuentre subestimado puesto que en el proceso de ensamblaje muchos de los elementos repetitivos han sido colapsados en "contigs" de pequeño tamaño y gran cobertura (Figura 40 C) generando un ensamblaje fraccionado. Incluso así, este ensamblaje presentó un espacio génico del 91 % según la herramienta CEGMA (97 % parciales). Además, este espacio génico se encontró ubicado en el 25 % de todos los "contigs" ensamblados, por lo que, aun estando fraccionado el ensamblaje debido a las zonas repetitivas, es muy probable que los "contigs"/"scaffolds" de mayor tamaño ensamblados constituyan el genoma completo.

*Estimación del tamaño nuclear mediante Real- Time PCR*

El tamaño del genoma nuclear de *Trebouxia* sp. TR9 se ha inferido con la técnica basada en PCR en tiempo real (RT-PCR) siguiendo la metodología de Armaleo & May (2009). La estimación del tamaño obtenido con esta técnica se encontró entorno a las 40-50 Mb. Este dato es similar al tamaño total del ensamblaje realizado (59 Mb), aunque es muy posible que sea una sub-estimación del tamaño nuclear total ya que en las secuencias ensambladas se observó que las zonas repetidas se habían colapsado en "contigs" de pequeño tamaño y alta cobertura, por lo que el tamaño del genoma nuclear de *Trebouxia* sp. TR9 ha de ser mayor. Además, en el trabajo de Armaleo & May (2009), la estimación del tamaño nuclear de *Asterochloris* sp. fue de 106,7 Mb, mientras que el tamaño ensamblado (http://genomeportal.jgi-psf.org/Astpho1/Astpho1.info.html) fue cercano a la mitad de este número (56,1 Mb). Esta técnica tiene el inconveniente de que usa medidas de concentración de ADN para las diferentes PCRs y RT-PCRs, y por ello, la concentración real del genoma nuclear se puede ver enmascarada por la presencia de ácidos nucleicos provenientes

de los genomas mitocondrial y cloroplástico (Armaleo, comunicación personal). Además, la alta sensibilidad de la RT-PCR puede verse comprometida por el error de pipeteo de los reactivos, con lo que se arrastraría un segundo error de precisión que comprometería la curva estándar y por consiguiente los cálculos realizados para la estimación.

El número de copias del ARN ribosomal nuclear calculado con la fórmula de Wang *et al.* (2011), es posible que también haya sido subestimado ya que se identificó este gen en un "contig" con cobertura alrededor del 70x (Node27_69900nt_73.58x) y en 2 "contigs" de alta cobertura y baja longitud (Node5_1980nt_1509.41x y Node915_5947nt_1543.45x). La elevada cobertura de estos últimos indica que son repeticiones colapsadas de este gen y que, realmente, podría encontrarse en cerca de 21 zonas diferentes del genoma.

Estos resultados muestran que la secuenciación masiva de ácidos nucleicos es más precisa, aunque presenten problemas de ensamblaje debido a las repeticiones y errores de secuenciación, que las técnicas utilizadas con anterioridad basadas en RT-PCR para la estimación del tamaño genómico o del número de repeticiones de un gen en dicho genoma.

*Anotación del genoma nuclear*

El genoma nuclear de *Trebouxia* sp. TR9 generado en esta tesis ha sido anotado usando factores predictivos de genes “ab initio” y de similitud con proteínas de otros organismos. Se han identificado un total de 9.499 genes, cantidad similar al de las otras algas trebouxiofíceas secuenciadas hasta el momento. El contenido en Guanina y Citosina ( %GC ) de los exones de los modelos génicos identificados es mayor que la media total de %GC del genoma de *Trebouxia* sp. TR9, como sucede en *Coccomyxa* y *Chlorella.* El número medio de exones por gen es también similar entre estas especies y otras clorofíceas secuenciadas (*Chlamydomonas reinhardtii* y *Volvox carteri*). Además, la composición de aminoácidos de los modelos proteicos inferidos mantienen una gran similitud con los de los diferentes organismos de las Viridiplantae. Todos estos indicios apuntan a que el genoma de *Trebouxia* sp. TR9 contiene proteomas estructuralmente y funcionalmente similares a los de otras algas de vida libre, y que no se han visto reducidos en número debido a los procesos de simbiosis liquénica, como sí que se ha producido en otros organismos simbiontes (Moya *et al.* , 2009). De hecho, el análisis de enriquecimiento de los términos GO muestran que en este alga, debido a su condición de productor primario en simbiosis liquénicas, la zona de contacto con el micobionte (la pared celular), las funciones moleculares de transferasas e hidrolasas, el metabolismo relacionado con el nitrógeno y los procesos de biosíntesis, han enriquecido su genoma. Este hecho, puede ser debido al modo de vida de este alga, ya que se encarga de

biosintetizar azúcares para el holobionte, y en especial para el micobionte, que es el mayor sumidero de carbono de este microecosistema. Palmqvist *et al.* (2002) propusieron que el micobionte es capaz de controlar la cantidad de carbohidratos disponibles en el holobionte regulando el tamaño de las poblaciones de fotobiontes, sus datos mostraron que la respiración del liquen aumentaba en relación con la cantidad de nitrógeno del talo, directa o indirectamente debido a la correlación entre el nitrógeno, la respiración y la cantidad de clorofila. Además, el metabolismo del nitrógeno puede ser una de los mecanismos de comunicación entre fotobiontes y micobiontes para la protección de los talos frente a la anhidrobiosis y la explosión de ROS en los primeros estadios de la rehidratación. Uno de las moléculas de señalización que parecen estar involucrados en este proceso es el óxido nítrico (NO), como apuntan los estudios realizados por nuestro grupo (Catalá *et al.* , 2010, 2013).

La anotación realizada es una primera aproximación al contenido real de los genes nucleares de *Trebouxia* sp. TR9, puesto que la metodología seguida tiene la limitación de la falta de evidencia experimental en forma de ARN. De todas las evidencias necesarias para la anotación de genomas, los datos de ARN son los que tienen el mayor potencial para aumentar la precisión de dicha anotación. Así, será necesario en un futuro, la obtención de este tipo de datos experimentales para poder delimitar de forma más adecuada los exones y sitios de "splice", transcritos alternativos para cada gen, RNAs no codificantes, zonas reguladoras y nuevos genes que AUGUSTUS ha dado como falsos negativos. Aún así, el primer boceto del genoma de *Trebouxia* sp. TR9 y su primera anotación serán útiles para la comunidad científica al ser el primer genoma de este género de alga simbionte de líquenes que se ha secuenciado y anotado en su casi totalidad.

*Dominios proteicos PFAM presentes en Trebouxia sp. TR9*

La comparación del porcentaje de motivos pertenecientes a familias de la base de datos PFAM presentes en los modelos proteicos de *Trebouxia* sp. TR9, es concordante con las proporciones de estas familias en otras especies de plantas. Este dato corrobora que el proceso de inferencia de genes, que se ha realizado para anotar el genoma de *Trebouxia* sp. TR9, es consistente. Una de las formas para averiguar la calidad de la anotación realizada en un genoma ensamblado "de novo", es la cuantificación del tanto por ciento de anotaciones que codifican proteínas con dominios conocidos (Yandell & Ence, 2012). Cuando se compararon los modelos PFAM compartidos por las algas de la clase Trebouxiophyceae (*Asterochloris* sp., *Chlorella variabilis*, *Coccomyxa subellipsoidea* y *Trebouxia* sp. TR9) y *Chlamydomonas reindhartii* (Chlorophyceae), aparecieron 59 dominios que pueden estar relacionadas con la adopción del modo de vida simbiótico de estos cuatro géneros de las trebouxiofíceas.

En el trabajo de Blanc *et al.* (2010) observaron que solo 27 familias proteicas eran propias de las algas verdes secuenciadas hasta ese momento. El nuevo genoma de *Trebouxia* sp. TR9 junto con los de *Asterochloris* sp. y *Coccomyxa subellipsoidea* han disminuido este número hasta 19. Conforme aparecen nuevos genomas de algas verdes, se constata que son pocas las familias proteicas que estaban ya en el ancestro común de Viridiplantae y que se han perdido en la vía evolutiva de la multicelularidad y la vida terrestre de Streptophyta.

Las 6 familias proteicas propias de los ficobiontes liquénicos *Trebouxia* sp. TR9 y *Asterochloris* sp. están implicadas en la regulación del ciclo celular, el metabolismo del nitrilo y algunas relacionadas con virus. Las primeras podrían ser responsables del mantenimiento de las poblaciones de estas algas en los talos liquénicos, puesto que aparece un motivo de apoptosis y otro de regulación mediada por calcio. Los modelos proteicos donde aparece la familia proteica de la "nitrile hidratase" son novedades en las Viridiplantae. Hasta la publicación de Marron *et al.* (2012), se creía que solo estaba en procariotas, en el coanoflagelado *Monosiga brevicollis* y en el estramenópilo *Aureococcus anophagefferens*. En dicho trabajo las detectaron en un amplio espectro de eucariotas, pero no en las Archaeplastida. Además, encontraron en diversos eucariotas que las sub-unidades alfa y beta de procariotas se habían fusionado en un solo gen, lo que sugiere que la fusión de estos genes debía encontrarse en el ancestro común de eucariotas (Marron *et al.* , 2012). En el caso de *Trebouxia* y *Asterochloris*, ambas poseen un modelo proteico que contiene ambos dominios, como en el caso de los otros eucariotas, pero que además la sub-unidad beta aparece en otro modelo proteico diferente en cada una de estas algas. Este hallazgo podría contradecir el argumento de que ambas sub-unidades se uniesen en un solo gen en el ancestro de eucariotas puesto que ambas algas tienen la sub-unidad beta en un gen independiente. Aunque no se puede descartar que dicha sub-unidad sea una duplicación de esta parte del gen que contiene ambas sub-unidades y que generó este nuevo gen en el ancestro de *Trebouxia* y *Asterochloris*. Para poder resolver este tipo de cuestión, será necesario secuenciar más genomas de algas verdes y corroborar cuál de estas dos hipótesis es la real. La aparición de motivos de cápsidas de virus de eucariotas, junto al hecho de que las secuencias que contienen esta familia proteica tienen la estructura exón-intrón típica de genes eucariotas, abre de nuevo dos interpretaciones: (i) podría ser un virus que se integró en el genoma de ambas algas y que han dado lugar a elementos virales endógenos (EVES), éstos ya se están empezando a identificar en los genomas de algas y plantas (Chu *et al.* , 2014). Sin embargo, tan solo un pequeño número de estas estructuras virales integradas en los genomas han podido ser detectadas. (ii) también cabe la posibilidad de que realmente sea un virus integrado en sus genomas. La presencia de virus en algas y en especial las liquénicas es un campo que acaba

de empezar a estudiarse (Petrzik *et al.* , 2015). Gracias a la tecnología de secuenciación masiva se podrá avanzar en este campo para poder diferenciar si realmente son virus funcionales como apuntan Petrzik *et al.* (2015) o EVES como encontraron Chu *et al.* (2014).

## 4.5 PROTEÍNAS IMPLICADAS EN EL METABOLISMO DEL CARBONO EN *trebouxia* SP. TR9

### 4.5.1 *Resultados*

*Proteínas implicadas en el transporte electrónico fotosintético*

Generalmente se acepta que los hongos liquénicos obtienen carbohidratos de los fotobiontes, es por ello que han de tener procesos de producción de energía eficientes para mantener al holobionte, como es el caso de la microalga *Trebouxia* sp. TR9. En la Figura 51, se representa el transporte electrónico fotosintético con los cuatro complejos proteínicos que llevan a cabo la primera fase de la fotosíntesis. En *Trebouxia* sp. TR9 al igual que en el resto de organismos fotosintéticos eucariotas, parte de estas proteínas están codificadas en el núcleo y parte en el cloroplasto. Los genes que codifican proteínas del fotosistema I *psaA*, *B*, *C, I* y *M* están codificados en el cloroplasto mientras que los genes *psaD, E, F, G, H, K, L, N* y *O* están codificados en el núcleo. Igualmente, los genes que codifican proteínas del fotosistema II se localizan tanto en el genoma cloroplástico como en el nuclear. Los genes *psbA, D, C, B, E, F, L, K, M, H, I, T* y *Z* están codificados en el cloroplasto mientras que los genes *psbO, P, Q, R, S, 27* y *28* están codificados en el núcleo. La mayoría de las subunidades del complejo citocromo b6/f están codificadas en el genoma cloroplástico por los genes *petA, B, D, G* y *L*. Solamente el gen *petC* se ha encontrado en el genoma nuclear. Los genes *petE, petF* y *petH* que codifican la Plastocianina (PC), la Ferredoxina (Fd) y Ferredoxin-NADP(+) oxydoreductasa (FNR), repectivamente, están codificados en el núcleo. Asimismo, las subunidades del complejo ATPasa de *Trebouxia* sp. TR9 están codificadas tanto en el genoma cloroplástico como en el nuclear. Los genes *atpA, B, E* que codifican las subunidades alfa, beta y épsilon del componente CF1, respectivamente, así como por los genes *atpH* e *I* que codifican las subunidades c y a del componente CFo, respectivamente, se localizan en el genoma cloroplástico. Sin embargo, los genes *atpG* y el *atpF* que codifican la sub-unidad gamma del componente CF1 y b del componente CFo, respectivamente, están codificados en el núcleo. Como puede observarse, la mayoría de las proteínas que soportan el trasporte electrónico fotosintético y la fotofosforilación están codificadas por genes ubicados en el genoma cloroplástico. Finalmente, hay algunos genes como *psaJ* y *X* del fotosistema I; *psbJ, U, V, W, X, Y Y* y *psb28-2* del fotosistema II; *petM* y *N* del complejo citocromo b6/f; y *petJ* que codifica el citocromo c6, que no se han encontrado.

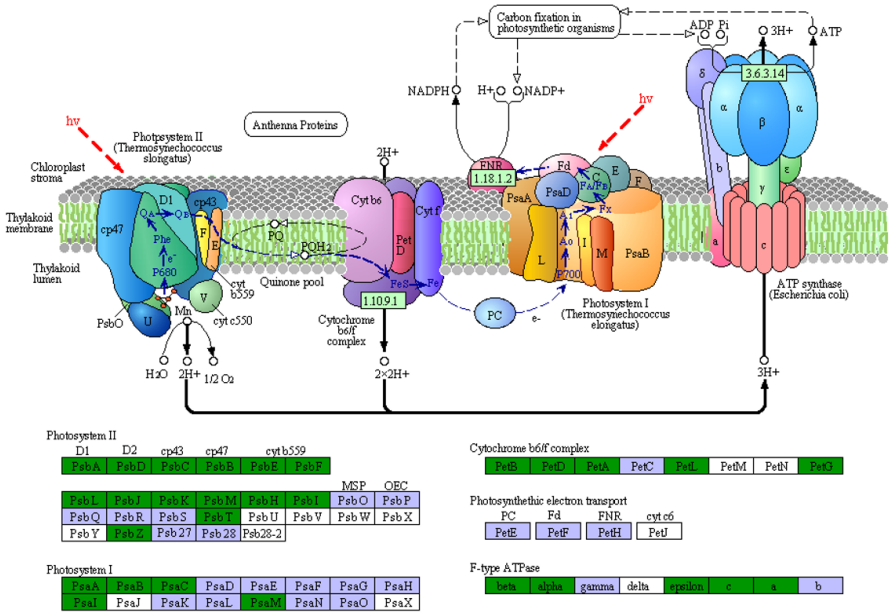



Figura 51: **Representación esquemática del transporte electrónico fotosintético adaptado del KEGG.** Las diferentes cajas son los genes de cada sub-unidad, el color de fondo rojo, verde y azul denota que el gen se encuentra en el genoma de la mitocondria, el cloroplasto o el núcleo respectivamente, las cajas de fondo blanco son genes que no están presentes en la anotación.

*Proteínas implicadas en la eficiencia de fijación de $CO_2$*

La actividad carboxilasa de la RuBisCO implica la unión de la enzima con Ribulosa 1,5-Bifostafo en sus centros activos, tras la aceptación del $CO_2$ esta actividad libera dos moléculas de 3-Fosfoglicerato, sin embargo, la actividad oxigenasa de la RuBisCO conlleva a la unión de $O_2$ en lugar del $CO_2$ produciendo Fosfoglicolato y 3-Fosfoglicolato. Esta actividad oxigenasa es el comienzo del proceso llamado fotorrespiración. En relación a la este proceso, se han encontrado en *Trebouxia* sp. TR9 secuencias similares a los componentes de este proceso metabólico (Figura 52 A). Algunas de ellas son: fosfoglicolato fosfatasa (EC 3.1.3.18), glicolato oxidasa (EC 1.1.3.15), malato sintasa (EC 2.3.3.9), malato deshidrogenasa (EC 1.1.1.37) y glutamato-glyoxylato aminotransferasa (EC 2.1.2.1).

Asimismo, hemos encontrado secuencias de enzimas similares a las de los mecanismos de concentración de carbono (CCM) como las que participan en las rutas metabólicas de fijación de carbono de plantas C3 y C4 y/o CAM (Figura 52 B). Se han encontrado tres modelos proteicos relacionados con enzimas responsables de catalizar la conversión rápida de dióxido de carbono y agua a bicarbonato y protones. Dos de ellas son "gamma carbonic anhydrase-like mitochondrial-like" y una "alpha carbonic anhydrase". Además, se han encontrado enzimas relacionadas con CCM del tipo C4, como son la fosfoenol

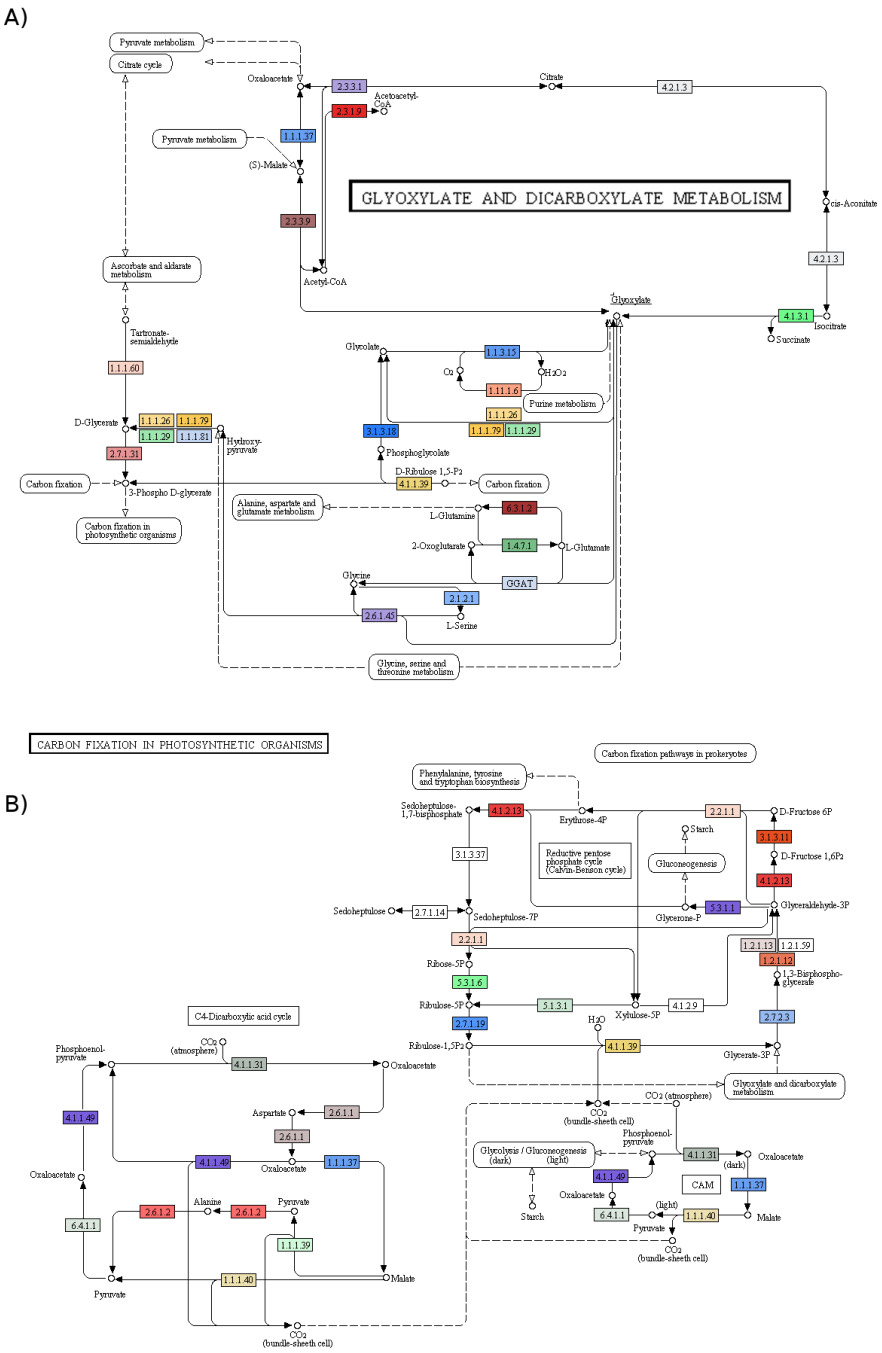



Figura 52: **Mapas metabólicos de los mecanismos de fotorrespiración (A) y concentración de carbono (B) de *Trebouxia* sp. TR9 (Adaptados del KEGG).** Las cajas de colores muestran las actividades enzimáticas anotadas del proteoma teórico de *Trebouxia* sp. TR9

piruvato carboxilasa (PEPC, EC 4.1.1.31), la malato deshidrogenasa (enzima málico, NADP-ME, EC 1.1.1.37) y la fosfoenol piruvato carboxiquinasa (PEPCK, EC 4.1.1.32). Además, se han encontrado posibles transportadores de ácidos de cuatro carbonos como la aspartato aminotransferasa (EC 2.6.1.1), la alanina aminotransferasa (EC 2.6.1.2) junto con secuencias similares a translocadores oxoglutarato-malato cloroplásticos que pueden facilitar el transporte de ácidos $C_4$ dicarboxílicos entre el cloroplasto y otros compartimentos subcelulares. Todos estos resultados sugieren que *Trebouxia* sp. TR9 posee dos mecanismos de concentración de carbono, unos de tipo $C_3$ junto a mecanismos de tipo $C_4$/CAM.

*Enzimas relacionadas con el metabolismo de carbohidratos*

Con el objeto de estudiar las proteínas hipotéticas que podrían ser enzimas relacionadas con el metabolismo de carbohidratos en *Trebouxia* sp. TR9 y otras algas de la clase Trebouxiophyceae (*Asterochloris* sp., *Coccomyxa subellipsoidea* y *Chlorella variabilis*), se utilizó la herramienta de anotación CAZymes Analysis Toolkit (`http://mothra.ornl.gov/cgi-bin/cat/cat.cgi`). Este programa utiliza la base de datos Carbohydrate Active enZimes (CAZy) para anotar las familias de proteínas pertenecientes a diferentes categorías: Actividades Auxiliares ("AA"), Módulos de Unión a Carbohidratos ("CBM"), Esterasa Carbónica ("CE"), Hidrolasas de Glucósidos (GH), Transferasas de Glucósidos ("GT") y Liasas de Polisacáridos ("PL") junto con los diferentes "clanes" que componen cada una de estas familias. La información recogida en esta base de datos diferencia las proteínas relacionadas con el metabolismo de carbohidratos mediante dos algoritmos diferentes, el primero se basa en la búsqueda de secuencias ortólogas utilizando el algoritmo BLAST y el segundo utiliza los motivos proteicos PFAM presentes en las secuencias analizadas y su interconexión con la correspondiente familia CAZyme.

En el presente trabajo, se han identificado al menos uno de los "clanes" de las familias CAZy en 1.047 y 918 modelos proteicos de *Trebouxia* sp. TR9 mediante búsquedas por ortología o utilizando la información del PFAM, respectivamente (Tabla 17). En el conjunto de las cuatro especies de algas que se han analizado, entre el 11 y 12 % de las proteínas totales pertenecían al menos a un clan de las familias CAZy.

En la Tabla 18 se muestra el número total de motivos de cada familia CAZy en los genomas de algas de la clase Trebouxiophyceae. Los "clanes" "CE16", "GT54", "GT45", "CBM37", "GT92" y "GT93" de las familias CAZy se han encontrado exclusivamente en *Trebouxia* sp. TR9. Las búsquedas por motivos PFAM en los modelos proteicos de *Trebouxia* sp. TR9 (Tabla 18) mostraron que solo 12 "clanes" constituían el 50,23 % (87 motivos "CE11", 83 motivos "GT48", 80 motivos "GT4", 66 motivos "GT2", 50 motivos "CBM48", 31 motivos "GH18",

Cuadro 17: **Número de proteínas que presentan motivos CAZy y "clanes" únicos de familias CAZy en las algas de la clase Trebouxiophyceae estudiadas.** Para cada especie se muestran las identidades obtenidas con búsquedas PFAM / búsquedas por ortología / comunes a ambas búsquedas.

| Especie | Nº Proteínas | Nº clanes CAZy |
|---|---|---|
| *Asterochloris* sp. | 895 / 823 / 385 | 108 / 119 / 95 |
| *Coccomyxa subellipsoidea* | 1133 / 980 / 535 | 118 / 139 / 100 |
| *Chlorella variabilis* | 1091 / 1134 / 509 | 120 / 138 / 103 |
| *Trebouxia* sp. TR9 | 1047 / 918 / 440 | 113 / 128 / 98 |

Cuadro 18: **Número total de motivos de cada familia CAZy en las proteínas de las algas de la clase Trebouxiophyceae secuenciadas.** Para cada especie se muestran las identidades obtenidas con búsquedas PFAM / búsquedas por ortología.

| Especie | AA | CBM | CE | GH | GT | PL |
|---|---|---|---|---|---|---|
| *Asterochloris* sp. | 55 / 31 | 94 / 271 | 113 / 57 | 273 / 211 | 380 / 273 | 0 / 0 |
| *Coccomyxa subellipsoidea* | 62 / 42 | 120 / 312 | 126 / 76 | 326 / 302 | 537 / 378 | 0 / 9 |
| *Chlorella variabilis* | 61 / 44 | 136 / 483 | 127 / 159 | 307 / 284 | 492 / 417 | 0 / 2 |
| *Trebouxia* sp. TR9 | 58 / 37 | 97 / 306 | 126 / 77 | 313 / 275 | 491 / 354 | 0 / 1 |

29 motivos “GT90”, 27 motivos “GT41”, 25 motivos “GT34”, motivos 23 “AA3”, 22 motivos “GH65” y 22 motivos “GH72”). En cambio, cuando se utilizaron búsquedas basadas en ortología el 50,71 % eran 13 "clanes" distintos a los anteriores (76 motivos “CBM20”, 72 motivos “GH18”, 61 motivos “GT2”, 53 motivos “CBM57”, 37 motivos “CE11”, 33 motivos “GT41”, 31 motivos “CBM48”, 30 motivos “GH38”, 29 motivos “CBM2”, 29 motivos “GT31”, 29 motivos “GT4”, 27 motivos “GT1” y 26 motivos “CBM50”).

Ante la discordancia de familias, "clanes" y modelos proteicos de *Trebouxia* sp. TR9 identificados con cada tipo de búsqueda, se optó por estudiar en detalle tan solo los modelos proteicos de *Trebouxia* sp. TR9 encontrados con ambos métodos. El número de modelos proteicos de *Trebouxia* sp. TR9 que contenían al menos un clan CAZy en ambas búsquedas fue de 440, encontrándose hasta 98 "clanes" de familias CAZy (Tabla 17). Estos "clanes" se repartían en 8 tipos de la familia “AA”, 10 de “CBM”, 5 de “CE”, 38 de “GH” y 38 de “GT”. Los 15 "clanes" de la Tabla 19 suponen el 50,32 % de todos los detectados. El “clan” “GT2”, localizado en 49 proteínas, es el que mayor representación tiene en los modelos proteicos de *Trebouxia* sp. TR9 (Tabla 19). Este “clan” fue el que mayor diversidad de actividades enzimáticas (“EC”) mostró seguido del “GT1” con 48 y 12, respectivamente. En cambio, los modelos proteicos de *Trebouxia* sp. TR9 que contenían

al clan "GH27" no se asociaron con ningún "EC", siendo anotados como proteínas relacionadas con acuaporinas.

Finalmente, con ambos métodos de búsqueda, se ha podido mejorar la anotación de 28 proteínas de *Trebouxia* sp. TR9 que no fueron anotados con BLAST2GO. Estas proteínas presentaron los motivos de los "clanes" CAZyme "AA2", "AA6", "AA10", "CBM13", "CBM20", "CBM35", "CBM50", "CBM48", "CE2", "GH1", "GT4", "GH23", "GH38", "GT8", "GT51", "GT61" y "GT90" y por tanto pueden estar relacionadas con el metabolismo de carbohidratos.

*Proteínas implicadas en la cadena respiratoria mitocondrial y fosforilación oxidativa*

En las plantas, el equilibrio de la respiración mitocondrial con los procesos fotosintéticos, determina la tasa de acumulación de biomasa. Las etapas finales del conjunto de reacciones que comprenden la respiración de plantas (Glucolisis, oxidación de pentosas fosfato, beta-oxidación de ácidos grasos, ciclo de ácidos tricarboxílicos) producen el poder reductor en forma de NADH y succinato para ser utilizados por la cadena de transporte electrónico mitocondrial y así mediante la fosforilación oxidativa convertir ADP en ATP para su posterior uso en otras reacciones celulares.

Las proteínas que intervienen en la cadena de transporte electrónico mitocondrial, están codificadas o bien en el núcleo o en el genoma mitocondrial. En el caso de *Trebouxia* sp. TR9, las proteínas que intervienen en la cadena de transporte electrónico mitocondrial y la fosforilación oxidativa (Figura 53), el complejo I NADH deshidrogenasa es en el que posee una mayor proporción de proteínas codificadas por genes que han permanecido en el genoma mitocondrial con 9 de 16 genes, seguido del complejo III Citocromo $b/c_1$ y del complejo V ATP sintasa con 3 de 6 y 4 de 9 genes, respectivamente.

### 4.5.2 *Discusión*

*Proteínas implicadas en el transporte electrónico fotosintético*

Las algas verdes, al igual que las plantas, obtienen gran parte de su materia y energía de la fotosíntesis. Este proceso biológico complejo implica la absorción de energía luminosa por biomoléculas fotosensibles que es transformada en una forma de energía bioquímica estable. La primera fase de la fotosíntesis es un proceso de conversión de energía luminosa en energía electroquímica. Como producto de reacciones de oxidación-reducción de proteínas asociadas a los tilacoides de los cloroplastos que generan un transporte de electrones y un bombeo de protones hacia el lumen tilacoidal, se obtiene en último término dos moléculas estables: NADPH y ATP. Estas biomo-

Cuadro 19: **Familias CAZyme más representadas en los modelos proteicos del genoma nuclear de *Trebouxia* sp. TR9 obtenidos con búsquedas basadas en ortología y por motivos PFAM.** Se indican los números enzimáticos (EC) presentes en los modelos proteicos de *Trebouxia* sp. TR9 que contenían dicho clan de la familia CAZy.

| Clan CAZy | Nº Proteínas | Descripción proteína y números enzimáticos (EC) asociados |
|---|---|---|
| GT2 | 49 | 3-oxoacyl-[acyl-carrier protein] (EC 2.3.1.41); 3-oxoacyl-[acyl-carrier-protein] synthase (EC 2.3.1.85); 4-aminobutyrate aminotransferase (EC 2.6.1.19 y EC 2.6.1.11); acetylornithine aminotransferase (EC 2.6.1.19 y EC 2.6.1.11); [Acyl-carrier-protein] S-malonyltransferase (EC 2.3.1.39); beta-ketoacyl synthase (EC 2.8.2.20, EC 1.1.1.36, EC 1.2.1.31, EC 1.1.1.1, EC 2.3.1.86, EC 6.2.1.3, EC 1.1.1.9, EC 2.3.1.39, EC 2.3.1.94, EC 1.6.5.5, EC 2.3.1.16 y EC 2.3.1.161); beta-ketoacyl synthase (EC:2.8.2.20, EC6.2.1.20, EC1.1.1.36, EC1.1.1.14, EC1.2.1.31, EC1.1.1.1, EC2.3.1.86, EC6.2.1.3, EC1.1.1.9, EC2.3.1.39, EC2.3.1.94, EC1.6.5.5, EC2.3.1.16 y EC2.3.1.161); beta-ketoacyl synthase (EC 6.2.1.3 y EC 2.3.1.94); chemotaxis protein (EC 2.7.3); cysteine--trna cytoplasmic-like isoform x2 (EC 6.1.1.16); cysteine--trna ligase-like (EC 6.1.1.16); dolichol-phosphate mannosyltransferase subunit 1 (EC 2.4.1.83 y EC 2.4.1.109); gamma-tocopherol methyltransferase (EC 2.1.1.95); glutamate-1-semialdehyde -aminomutase (EC 5.4.3.8); guanylate cyclase (EC 2.7.3); histidine kinase (EC 2.7.3); malonyl-coenzyme A-acyl carrier protein mitochondrial-like (EC 2.3.1.39); phosphoethanolamine n-methyltransferase 1-like (EC 2.1.1 y EC 2.1.1.103); polyketide synthase (EC 2.3.1. EC 2.3.1.94 y EC 1.1.1.100); probable rhamnose biosynthetic enzyme 1-like (EC 4.2.1.46, EC 1.1.1.133 y EC 4.2.1.74); probable rhamnose biosynthetic enzyme 1-like isoform x1 (EC:4.2.1.46, EC 4.2.1.76 y EC 1.1.1.133); udp-glucose 4-epimerase gepi48-like (EC 5.1.3.2 y EC 5.1.3.5); udp-glucose dehydrogenase (EC 1.1.1.22); uncharacterized hydrolase yor131c-like (EC 3.1.3.18) |
| CE11 | 37 | aminoacyl-histidine dipeptidase (EC 3.5.1); atp-dependent rna helicase ddx1-like (EC 2.3.1 y EC 3.4.24); dead-box atp-dependent rna helicase 38-like (EC 2.7.7); dead-box atp-dependent rna helicase 7-like (EC 2.7.11); dead-box atp-dependent rna helicase mitochondrial (EC 2.7.7); peptidyl-prolyl cis-trans isomerase pasticcino1-like (EC 5.2.1.8); rna helicase drh1 (EC 6.1.1.2) |
| CBM48 | 20 | 1,4-alpha-glucan branching enzyme (EC 2.4.1.18); isoamylase chloroplastic-like (EC 3.2.1.1 y EC 3.2.1.68); probable e3 ubiquitin-protein ligase herc4 isoform x2 (EC 6.3.2.19); probable monogalactosyldiacylglycerol chloroplastic-like (EC 2.4.1.46); pullulanase chloroplastic-like (EC 3.2.1.142, EC 3.2.1.41 y EC 3.2.1.1); starch branching enzyme (EC 2.4.1.18); thiamin pyrophosphokinase 2 (EC 2.7.6.2); ubiquitin-protein ligase e3a isoform x1 (EC 6.3.2.19) |
| GT1 | 17 | Adenosinetriphosphatase (EC 3.6.1.3); Alpha,alpha-trehalose-phosphate synthase (UDP-forming) (EC 2.4.1.15); alpha- -mannosyl-glycoprotein 2-beta-n-acetylglucosaminyltransferase-like (EC 2.4.1.101 y EC 2.4.1.94); diacylglycerol o-acyltransferase 2-like (EC 2.3.1.20); diacylglycerol o-acyltransferase (EC 2.3.1.51); glycogen phosphorylase 1-like (EC 2.7.1.11); glycosyltransferase family 15 protein (EC 2.4.1.131); granule-bound starch synthase chloroplastic amyloplastic-like (EC 2.4.1.21 y EC 4.3.3.2); soluble starch synthase chloroplastic amyloplastic-like (EC 2.4.1.21); starch synthase chloroplastic amyloplastic-like (EC 2.4.1.21); sucrose synthase (EC 2.4.1.13); udp-glucose:glycoprotein glucosyltransferase (EC 2.4.1) |
| GT4 | 14 | 3-beta-hydroxysteroid-delta -isomerase (EC 5.3.3.3 y EC 5.3.3.5); 3-beta-hydroxysteroid-delta -isomerase (EC 5.3.3.3 y EC 5.3.3.5); gamma-tocopherol methyltransferase (EC EC:2.1.1.95); sucrose synthase (EC 2.4.1.13) |
| GH1 | 13 | atpase (EC 3.6.3.1); diacylglycerol o-acyltransferase 2-like (EC 2.3.1.20); diacylglycerol o-acyltransferase 2-like (EC 2.3.1.20); diacylglycerol o-acyltransferase (EC 2.3.1.51); udp-glucuronosyltransferase 2b31-like (EC 2.3.1.51) |
| GT66 | 13 | peptidyl-prolyl cis-trans cyclophilin-type (EC 5.2.1.8); peptidyl-prolyl cis-trans isomerase cwc27 homolog (EC 5.2.1.8); peptidyl-prolyl cis-trans isomerase cyp20-1-like (EC 5.2.1.8); peptidyl-prolyl cis-trans isomerase cyp20- chloroplastic-like (EC 5.2.1.8); peptidyl-prolyl cis-trans isomerase cyp20-chloroplastic-like (EC 5.2.1.8); peptidyl-prolyl cis-trans isomerase d (EC 5.2.1.8); peptidyl-prolyl cis-trans isomerase (EC 5.2.1.8); peptidyl-prolyl cis-trans isomerase-like 1 (EC 5.2.1.8); peptidyl-prolyl cis-trans isomerase-like 2-like (EC 2.4.1.119); peptidyl-prolyl cis-trans isomerase-like 3-like (EC 5.2.1.8); peptidyl-prolyl cis-trans isomerase-like 4-like (EC 5.2.1.8); peptidylprolyl isomerase domain and wd repeat-containing protein 1-like (EC 5.2.1.8) |
| GH18 | 11 | 2-cys peroxiredoxin chloroplastic-like (EC 1.11.1.7 y EC 1.11.1.15) y peroxiredoxin-6 (EC 1.11.1.15, EC 1.11.1.7 y EC 3.1.1.4); glutathione peroxidase (EC 1.11.1.7 y EC 1.11.1.15); glutathione peroxidase (EC 1.11.1.9 y EC 1.11.1); peroxiredoxin-1 isoform x2 (EC 1.11.1); protein rmd5 homolog a-like (EC 3.2.1.14); valine--trna ligase-like (EC 6.1.1.9); valine--trna mitochondrial (EC 6.1.1.9) |
| GT48 | 11 | d-galactose transporter (EC 1.3.1.74); rna-directed dna polymerase from mobile element jockey-like (EC 2.7.7.49) |
| GH38 | 11 | lysosomal alpha-mannosidase (EC 3.2.1.24); m7 diphosphatase (EC 3.6.1.30) |
| GT31 | 10 | beta -galactosyltransferase 7-like (EC 2.4.1); coatomer subunit gamma-like (EC 2.4.1); hypothetical protein COCSUDRAFT_48261 (EC 2.4.1); hypothetical protein COCSUDRAFT_64264 (EC 2.4.1); probable beta- -galactosyltransferase 14-like (EC 2.4.1); probable beta- -galactosyltransferase 19-like (EC 2.4.1) |
| GT34 | 10 | carotene epsilon-chloroplastic-like (EC 1.14.99); cytochrome p450 chloroplastic-like (EC 1.14.14.1) |
| GT41 | 10 | peptidyl-prolyl cis-trans isomerase pasticcino1-like (EC 5.2.1.8); probable udp-n-acetylglucosamine--peptide n-acetylglucosaminyltransferase spindly-like isoform x1 (EC 2.4.1.94); tetratricopeptide repeat protein 13-like (EC 6.3.2.12 y EC 6.3.2.17); uncharacterized mitochondrial protein g00810-like (EC 2.3.1) |
| GH27 | 9 | aquaporin pip2-2-like; aquaporin pip2-7-like; probable aquaporin tip3-2-like |
| GH92 | 9 | 2-deoxy-d-gluconate 3-dehydrogenase (EC 1.1.1.100); 3-oxoacyl-[acyl-carrier-protein] reductase (EC 1.1.1.100); cytochrome c biogenesis protein (EC 1.1.1.30); nad -binding protein (EC 2.1.1); nad -binding protein (EC 5.2.1.8) |

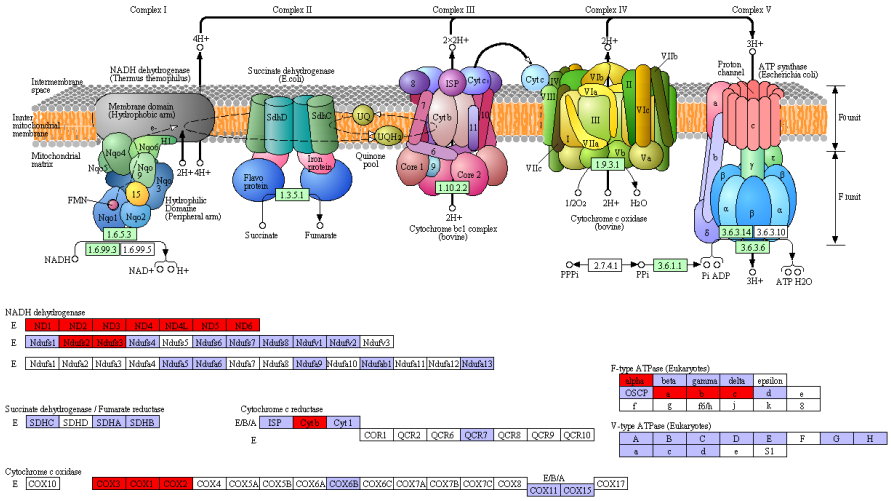



Figura 53: **Mapa metabólico de la fosforilación oxidativa de *Trebouxia* sp. TR9 (Adaptados del KEGG).** Las diferentes cajas son los genes de cada sub-unidad, el color de fondo rojo o azul denota que el gen se encuentra en el genoma de la mitocondria o del núcleo respectivamente, las cajas de fondo blanco son genes que no están presentes en la anotación. A=Arquea, B=Bacteria y E=Eucariota.

léculas proporcionan poder el reductor (NADPH) y la energía (ATP) necesarios para la fijación y asimilación efectiva del $CO_2$.

Los cloroplastos son orgánulos que provienen de la incorporación de una cianobacteria por endosimbiosis y contienen mucho menos ADN que sus relativos contemporáneos. Esta pérdida de ADN es una consecuencia de la redistribución del material genético entre el cloroplasto y el núcleo a lo largo de la evolución (Kleine *et al.* , 2009). Muchos de los procesos bioenergéticos son controlados por el núcleo al ser éste el que provee de proteínas para el correcto funcionamiento de las cadenas de transporte electrónico. El uso de los datos proporcionados por la base de datos KEGG, ha permitido comparar el contenido y la procedencia de los genes de la cadena de transporte fotosintético de *Trebouxia* sp. TR9 con el de *Arabidopsis thaliana*. Se ha observado un patrón muy similar de distribución de genes entre el genoma cloroplástico y el nuclear con la excepción del gen *atpF*, que codifica la sub-unidad b del componente CFo de la ATPasa del cloroplasto. En *Trebouxia* sp. TR9 este gen parece estar codificado en el núcleo a diferencia de lo que sucede en *Arabidopsis*, que está codificado en el cloroplasto. De igual forma, tanto en *Trebouxia* sp. TR9 como en las algas *Chlorella variabilis*, *Coccomyxa subellipsoidea*, *Chlamydomonas reindhartii* y en *Arabidopsis thaliana*, los genes *psbU*, *V*, *X* y *28-2* pertenecientes al fotosistema II, el gen *psaX* del fotosistema I y el gen *petM* del complejo citocromo b6/c no están presentes en ninguna de estos organismos. Lo que sugiere que estos genes, presentes en otros organismos fotosintéticos, se perdieron en el ancestro común de algas verdes y plantas terrestres. El gen *psbW* no está presente

en los genomas de *Chlorella variabilis* y *Trebouxia* sp. TR9, por lo que se debió de perder en la separación de estos géneros respecto a las otras trebouxiofíceas. El gen *petL* del complejo citocromo b6/c se ha mantenido en *Trebouxia* sp. TR9, mientras que en el resto de las algas anteriormente citadas se ha perdido. Por último, los genes del fotosistema II *psbY* y *psbJ*, el gen *petN* del complejo citocromo b6/c, y el gen de la sub-unidad delta de la ATPasa no están presentes en los genomas de *Trebouxia* sp. TR9. Estos datos abren una puerta a la investigación sobre cuáles son las sub-unidades y los mecanismos de control mínimos necesarios para que funcione adecuadamente el trasporte electrónico fotosintético, las diferentes lineas evolutivas que han aparecido a partir del primer evento de endosimbiosis y los genes que se han conservado/perdido durante la evolución entre otras.

*Proteínas implicadas en la eficiencia de fijación de $CO_2$*

La actividad oxigenasa de la RuBisCO es el comienzo de la fotorrespiración, que comprende una serie de reacciones llamadas ciclo $C_2$. El Fosfoglicolato formado en el cloroplasto es rápidamente hidrolizado a glicolato y derivados que son tóxicos. Esta actividad en *Trebouxia* sp. TR9 produce fosfoglicolato, que por acción de la fosfoglicolato fosfatasa se transforma en glicolato. La oxidación del glicolato a glioxilato es catalizada en *Trebouxia* sp. TR9 por la glicolato oxidasa típica de plantas traqueofitas. Este metabolito puede ser reciclado por dos vías diferentes:

i) El glioxilato puede ser condensado con acetil-CoA a malato por la malato sintasa y seguidamente a oxalacetato por la malato deshidrogenasa. Finalmente el oxalacetato puede seguir dos rutas diferentes, el metabolismo del piruvato o el ciclo de ácidos tricarboxílicos seguido del metabolismo del ascorbato y renovar el D-glicerato, el 3-fosfo D-glicerato y producir D-Ribulosa 1,5-$P_2$de nuevo por la actividad de la RuBisCO.

ii) Dos glioxilatos pueden ser convertidos de nuevo en 3-fosfo D-glicerato para comenzar el proceso fotosintético y producir D-Ribulosa 1,5-$P_2$, primero convirtiendo uno de ellos en glicina por la glutamato-glyoxylato aminotransferasa, ésta a L-serina con la cual el segundo glioxilato se convierte en hidroxipiruvato, el cual es sustrato para obtener el D-glicerato que es convertido en 3-fosfo D-glicerato.

Este proceso de fotorrespiración reduce el rendimiento energético en la asimilación de $CO_2$, pero es importante para contribuir a disipar el exceso de energía de la cadena de transporte electrónico fotosintética que puede formarse en condiciones de elevada iluminación y bajas concentraciones de $CO_2$ y así evitar el daño y estrés oxidativo debido a la creación de especies reactivas del oxígeno. Uno de las formas que algunas plantas tienen para evitar la actividad oxigenasa de la RuBisCO son los mecanismos de concentración de $CO_2$ (CCM). Estos mecanismos son comunes en las algas verdes que crecen en hábitats

acuáticos donde la baja difusión del CO2 limita la capacidad fotosintética por lo que, para evitar procesos de fotorrespiración o de estrés oxidativo, utilizan estos mecanismos para aumentar la concentración de carbono inorgánico ($CO_2$ y $HCO_{3-}$) entorno a la enzima RuBisCO y así tener una mejor eficiencia en la fijación del carbono atmosférico. En el caso de los talos liquénicos, al ser organismos poiquilohídricos, la cantidad de agua es la limitante para los procesos de fijación de carbono. En las algas verdes liquénicas analizadas por Palmqvist (2000) se detectaron CCMs, aunque los mecanismos genéticos eran desconocidos.

En *Trebouxia* sp. TR9 hemos encontrado secuencias similares a los CCM de plantas C3 junto a enzimas de rutas metabólicas de fijación de carbono del tipo C4 y/o CAM como son la fosfoenolpiruvato carboxilasa (PEPC), la malato deshidrogenasa (enzima málico, NADP-ME) o la fosfoenolpiruvato carboxiquinasa (PEPCK). Además, se han encontrado transportadores de ácidos de cuatro carbonos como la aspartato aminotransferasa, la alanina aminotransferasa y posibles translocadores oxoglutarato-malato cloroplásticos, todos ellos encargados del transporte de ácidos C4 en diferentes compartimentos celulares. Todos estos resultados sugieren que *Trebouxia* TR9 posee unos mecanismos de concentración de carbono de tipo C4/CAM similares a los de *Myrmecia* (Ouyang *et al.* , 2013) junto a los de otras algas verdes que presentan unos CCM basados en anhidrasas carbónicas como *Coccomyxa subellipsoidea* y *Chlamydomonas reindhartii* (Blanc *et al.* , 2012).

En los experimentos realizados por Cowan *et al.* (1992) con *Ramalina maciformis* y recopilados por Green *et al.* (2008), se pudo observar que el fotobionte del género *Trebouxia* alcanzaba diferentes valores para la discriminación de $^{13}C$ dependiendo del contenido de agua del liquen. Cuando el contenido de agua era del 50 %, la discriminación de $^{13}C$ presentaba valores parecidos a los de plantas C3 (-23.1 y -20.2 ante baja y alta irradiación, respectivamente). Cuando el contenido de agua era del 120 %, la discriminación isotópica de $^{13}C$ era más del tipo de plantas C4 (-19.8 y -5.4 ante baja y alta irradiación, respectivamente). Nuestro grupo ha empezado a realizar experimentos de discriminación de isótopos de carbono $^{13}C$ en las algas que coexisten en los talos de *Ramalina farinacea, Trebouxia* sp. TR9 y *Trebouxia jamesii* (Molins A., Conesa M.A. y Ribas-Carbó M., en progreso). Para ello, se calculó la discriminación de $^{13}C$ de estas algas crecidas en medio autotrófico (3NBBM). En el caso de *Trebouxia* sp. TR9 los valores obtenidos fueron similares a los de plantas C4 (-15,84). En el caso de *Trebouxia jamesii* los valores fueron más cercanos a los de plantas C3 (-21,7). Estos resultados podrían ser un buen inicio para tratar de entender los patrones de ficobiontes predominantes en los talos de *Ramalina farinacea* analizados por del Campo *et al.* (2013). En ese trabajo se observó que los talos procedentes de Canarias presentaban

predominancia de *Trebouxia* sp. TR9, mientras que en los talos de la Península Ibérica, el ficobionte más abundante era *Trebouxia jamesii.* Puesto que las condiciones meteorológicas de las zonas de Canarias donde se recolectaron los talos para la experimentación, son mucho más húmedas debido al régimen de nieblas, que en las zonas de la Península Ibérica analizadas. El hecho de que *Trebouxia* sp. TR9 se comporte de forma similar a plantas $C_4$, probablemente debido a los mecanismos de concentración de carbono revelados en el presente trabajo, puede suponer una ventaja en este tipo de ambientes con elevada humedad atmosférica y de alta irradiación, donde la difusión del $CO_2$ en el interior del talo será menor que en los talos crecidos en ambientes más xéricos de la Península Ibérica, donde *Trebouxia jamesii* es el fotobionte predominante. Los mecanismos de concentración de carbono de tipo $C_4$, en estos ambientes, implicarían una pérdida de eficiencia fotosintética debido al mayor gasto energético de la ruta $C_4$ en comparación con la ruta $C_3$.

De todas formas, será necesario continuar la investigación sobre los mecanismos de concentración de carbono para poder dilucidar si son de tipo $C_4$ o CAM, determinar los niveles de expresión de todas estas proteínas y las diferencias a nivel genómico, transcriptómico y proteico entre *Trebouxia* sp. TR9 y *Trebouxia jamesii,* para confirmar estos resultados e incluso poder generar las herramientas genéticas para aplicaciones biotecnológicas como, por ejemplo, la mejora de ciertos cultivos agrícolas.

*Enzimas relacionadas con el metabolismo de carbohidratos*

Los polisacáridos constituyen la clase más diversa de macromoléculas presentes en la naturaleza tanto en términos de estructura primaria, ya que cuentan con una amplia variedad de monómeros de azúcar que pueden ser unidos de diversas formas (lineal o ramificadamente) y por el gran número de residuos que se pueden unir lateralmente (por ejemplo, grupos sulfato, metilo, o acetilo). El número de modelos proteicos anotados en el genoma de *Trebouxia* sp. TR9 como el de "clanes" de familias CAZy son parecidos a los de las algas trebouxiofíceas secuenciadas (Tabla 17), siendo las algas simbiontes de líquenes *Asterochloris* sp. y *Trebouxia* sp. TR9 las que presentan un menor número de proteínas con motivos de las familias de CAZyme, aunque el tanto por ciento de las proteínas de las cuatro algas analizadas es constante, entorno al 12 %. Una característica común en estas algas es el bajo número de la familia de CAZyme perteneciente a las liasas que actúan sobre polisacáridos (“PL”) y un alto número de familias relacionadas con la hidrólisis y transferencia de glicósidos (“GH” y “GT”). Este bajo número de identificaciones de “clanes” de la familia “PL” en las algas de la clase Trebouxiophyceae analizadas puede ser debido a que la familia “PL” incluye un grupo de enzimas que escinden cadenas de polisacáridos que contienen ácido urónico a través de

un mecanismo de beta-eliminación para generar un residuo de ácido hexenurónico insaturado y un nuevo extremo reductor, el bajo número de identificaciones obtenidas se puede deber a que en la base de datos CAZy la familia de liasas de polisacáridos es frecuentemente poliespecífica (es decir, contienen enzimas que actúan sobre diferentes sustratos o que generan diferentes productos) y por tanto muchas de las liasas carbono-oxígeno que actúan sobre polisacáridos bajo el número enzimático EC 4.2.2.- han sido identificadas como parte del resto de familias CAZy (Lombard *et al.* , 2010). Otra posible explicación sería que como muchas enzimas de este tipo de familia muestran una gran variedad de tipos de plegado que sugiere que las "PL" han aparecido más de una vez durante la evolución (Lombard *et al.* , 2010) y en las algas analizadas en este trabajo no han aparecido. Esta segunda explicación es factible con los datos ofrecidos por la base de datos PlantCAZyme (`http://cys.bios.niu.edu/plantcazyme/`) (McGinn *et al.* , 2014), donde las plantas analizadas tan solo presentaron los "clanes" "PL1", "PL4", "PL5", "PL8" y "PL12" mientras que *Chlamydomonas reinhardtii* de las Chlorophyceae, tampoco presenta ninguno de estos "clanes" de la familia "PL". Estos resultados indicarían que la familia "PL" no ha aparecido durante la evolución de las algas de la división Chlorophyta. Los enzimas implicados en la biosíntesis de ramnosa están también muy representados en las identificaciones, lo que concuerda con las observaciones de que la pared de *Trebouxia* sp. TR9 es rica en ramnosa Casano *et al.* (2015).

Para identificar las familias CAZy en los modelos proteicos de *Trebouxia* sp. TR9 se han utilizado dos metodologías diferentes. Una basada en búsqueda de ortologías utilizando identificaciones recíprocas de BLAST (Reciprocal BLAST Hit, RBH), con las que se pueden cometer errores de tipo I (Falsos positivos) al identificar modelos proteicos que se parecen en una parte de la proteína donde no se encuentra la actividad relacionada con el metabolismo de carbohidratos. Además, tiene el problema de cometer un error de tipo II (Falsos negativos) con las proteínas parálogas de *Trebouxia* sp. TR9. La segunda metodología se basa en la interrelación de la base de datos CAZy con la de familias proteicas PFAM. En este caso, el posible problema reside en que si la familia CAZy no tiene correlación con familias del PFAM, se cometerá un error de tipo II. Es por ello que, de una forma conservadora, tan sólo los modelos proteicos de *Trebouxia* sp. TR9 que fueron identificados por ambos métodos se han utilizado para analizar los números de enzima ("EC") anotadas en ellos (Tabla 19).

Los modelos proteicos que contenían la mayoría de "clanes" CAZy se relacionan con las estrategias que *Trebouxia* sp. TR9 podría utilizar para evitar el estrés oxidativo debido a los procesos de deshidratación a los que se ve sometida en su medio natural (del Hoyo *et al.* , 2011; Casano *et al.* , 2011). Se han encontrado un gran número de actividades enzimáticas relacionadas con la síntesis lipídica para po-

der mantener y recomponer las membranas celulares o generar otras sustancias de reserva además del almidón. Otra de las características que se han encontrado para evitar este tipo de estrés es el control de la expresión proteica (helicasas o la síntesis de ARNt) y el rápido post-procesamiento de las proteínas mediante ubiquitinización, fosoforilación o metilación. Otras de las posibles estrategias de *Trebouxia* sp. TR9 para evitar el estrés oxidativo se pueden relacionar con compuestos basados en la acción del glutation, la peroxireducción, la disipación del exceso energético lumínico (carotenos) y/o la protección de los procesos de fotosíntesis-respiración (Citocromo-p 450), junto con el control del estado hídrico de las células (Aquaporinas).

Aunque la metodología utilizada ha sido conservadora y sería necesario analizar los modelos proteicos de *Trebouxia* sp. TR9 que han aparecido en cada uno de ellos por separado y que no son comunes a ambos tipos de búsquedas, los resultados concuerdan con los aspectos fisiológicos a los que está sometida esta alga en su hábitat natural. Además, estudios previos sobre el control del estrés oxidativo (Casano *et al.* , 2011; del Hoyo *et al.* , 2011; Álvarez *et al.* , 2012, 2014) o la composición de su pared celular (Casano *et al.* , 2015) en este taxón, parecen corroborar estos resultados.

*Proteínas implicadas en la cadena respiratoria mitocondrial y fosforilación oxidativa*

Las mitocondrias son orgánulos que provienen de la incorporación por endosimbiosis de una alfa-proteobacteria por un pre-eucariota. Estos orgánulos contienen mucho menos ADN que sus relativos bacterianos contemporáneos. Esta pérdida de ADN ha sido causada durante la evolución por la redistribución del material genético entre el núcleo y la mitocondria (Kleine *et al.* , 2009) junto a la pérdida de genes bacterianos por procesos evolutivos. El hecho de que se codifiquen muchas de las proteínas en los genomas de los orgánulos puede ser debido a que sus productos proteicos sean muy hidrofóbicos y no puedan ser importados al orgánulo desde el citosol, que dichos productos sean tóxicos cuando están presentes en el citosol u otros componentes celulares inapropiados o que estos genes se han mantenido en los genomas organulares para facilitar la rápida regulación de su expresión por el estado redox del orgánulo (Odintsova & Yurina, 2005).

Los genes que codifican las diferentes sub-unidades de los complejos de la cadena de trasporte electrónico y fosforilación oxidativa de *Trebouxia* sp. TR9, *Chlorella variabilis* y *Coccomyxa subellipsoidea*, se puede observar que el número de genes conservados del Complejo I de *Coccomyxa subellipsoidea* es mayor mientras que en *Chlorella variabilis* es menor que en *Trebouxia* sp. TR9 y que en las tres algas no esta presente la "NADH dehydrogenase [ubiquinone] flavoprotein 3", al igual que en el resto de plantas secuenciadas, por lo que es-

te gen se perdió en el ancestro común de plantas terrestres y algas verdes. En relación al complejo II succinato deshidrogenasa, *Coccomyxa subellipsoidea* es de las tres algas estudiadas la que tiene todas las sub-unidades mientras que en *Trebouxia* sp. TR9 falta el gen para la proteína "Succinate dehydrogenase [ubiquinone] cytochrome b small subunit" (SDHD) y en *Chlorella variabilis* falta el gen de la proteína "Succinate dehydrogenase subunit 3-2" (SDHC). En el complejo III citocromo b/c1 de *Trebouxia* sp. TR9 no han sido identificados los genes para las proteínas "ubiquinol-cytochrome c reductase subunit 6" y "ubiquinol-cytochrome c reductase subunit 9" (QCR6 y QCR9, respectivamente) que sí están presentes en los genomas de *Chlorella* y *Coccomyxa*. Además, en *Trebouxia* sp. TR9 tampoco se encontraron los genes para las proteínas COX5B, COX6A, COX10 y COX17 del complejo IV citocromo c oxidasa que en *Chlorella* y *Coccomyxa* sí que están presentes. Finalmente, en el complejo V ATP sintasa, en *Trebouxia* sp. no se encontraron los genes para las sub-unidades g y épsilon que en *Coccomyxa* y *Chlorella* están presentes, además, en *Chlorella* faltan las sub-unidades delta y d.

## 4.6 Análisis del exoproteoma de *trebouxia* sp. TR9 y *trebouxia jamesii*

### 4.6.1 *Resultados*

Como continuación al trabajo de Casano *et al.* (2015), se ha llevado a cabo la secuenciación de diferentes exoproteínas de pared de *Trebouxia* sp. TR9 y *T. jamesii* tras ser tratadas durante una semana con 0 y 100 µM de $Pb(NO_3)_2$. Se realizaron dos tipos de secuenciaciones proteómicas: los extractos totales de polipéptidos extracelulares de *Trebouxia* sp. TR9 y *T. jamesii* control fueron digeridos en suspensión líquida y secuenciados mediante cromatografía líquida acoplada a espectrometría de masas; y en segundo caso, se recortaron diferentes bandas de los polipéptidos separados en una electroforesis monodimensional en gel de poliacrilamida SDS-PAGE (Figura 54). En ambos casos, los polipéptidos (El total en el primer caso, o cada banda por separado en el segundo caso) fueron sometidos a digestión con tripsina, la cual generó un conjunto de péptidos, los cuales fueron separados y caracterizados por espectrometría de masas. Se espera obtener un espectro de masas (MS) característico y único denominado "Huella Peptídica" para cada polipéptido. La identificación de los péptidos trípticos (pertenecientes a una proteína particular) se realizó automáticamente mediante el sistema GPS (Global Protein Server) y usando como motor de búsqueda MASCOT (MatrixScience, UK) frente a dos base de datos de proteínas alternativas.

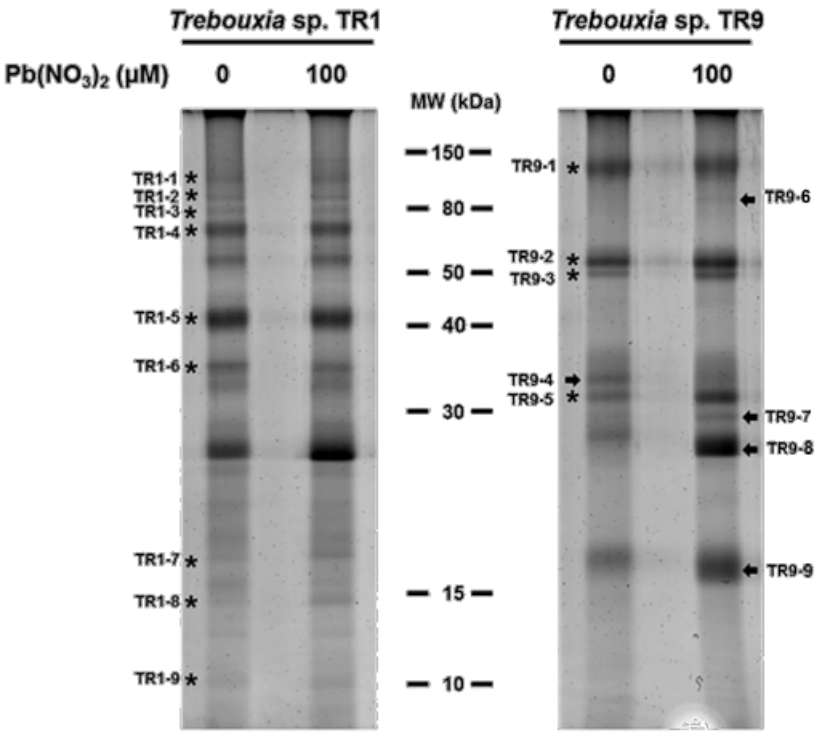



Figura 54: **Figura gel monodimensional SDS-PAGE de proteínas extracelulares de *Trebouxia jamesii* (TR1) y *Trebouxia* sp. TR9.** Se indican con un asterisco y con una flecha las bandas que se secuenciaron, en el primer caso son bandas que cuya abundancia cambió con el tratamiento, mientras que el segundo caso son bandas que aparecieron/desaparecieron según el tratamiento.

Cuadro 20: **Comparación de los exopéptidos detectados utilizando la base de datos del NCBI frente a los modelos proteicos de *Trebouxia* sp. TR9 de los tratamientos control de la digestión en líquido realizada.**

| | **Péptidos detectados** | **Max score** | **Min score** | **Media score** | **Max match** |
|---|---|---|---|---|---|
| TR9 Illumina | 81 | 293 | 20 | 60 | 27 |
| TR9 NCBI | 45 | 63 | 22 | 44 | 20 |
| *T. jamesii* Illumina | 129 | 899 | 20 | 62 | 52 |
| *T. jamesii* NCBI | 71 | 278 | 21 | 51 | 46 |

*Identificación de los péptidos extracelulares obtenidos por digestión en líquido*

En la Tabla 20 se comparan los péptidos obtenidos de la digestión en suspensión líquida e identificados empleando dos bases de datos diferentes: El NCBI-Viridiplantae (Noviembre 2014), y la anotación de los modelos proteicos de *Trebouxia* sp. TR9 obtenidos en esta Tesis. El número de péptidos, las puntuaciones (score) y los péptidos coincidentes para cada proteína de la base de datos (match), son mayores cuando se utilizó la base de datos de modelos proteicos de *Trebouxia* sp. TR9 frente a la base de datos de proteínas del NCBI, tanto para *T. jamesii* como para *Trebouxia* sp. TR9 (Tabla 20). En las identificaciones obtenidas de los péptidos contra la base de datos del NCBI, en el caso de *Trebouxia* sp. TR9 han aparecido como *rbcL* 12 de 45 identificaciones y en *T. jamesii* han sido 19 de 71 identificaciones. En las identificaciones realizadas con la base de datos de modelos proteicos de *Trebouxia* sp. TR9, en cada una de las especies, tan sólo una de las identificaciones ha correspondido a esta proteína, en *Trebouxia* sp. TR9 y en *T. jamesii* esta identificación poseía 27 y 52 péptidos coincidentes respectivamente (datos no mostrados). Esto corrobora, no solo que el número de identificaciones realizadas con la base de datos del NCBI es menor, sino que existe una gran redundancia en ellas debido a la poca cantidad de proteínas del género *Trebouxia* existentes en esta base de datos (tan sólo existen algunas proteínas de origen cloroplástico, mitocondrial, la actina y la proteína ribosomal L10). Es por ello que el trabajo de identificación de las exoproteínas, y por lo tanto el análisis de sus péptidos trípticos se llevó a cabo empleando la base de datos de los modelos proteicos de *Trebouxia* sp. TR9.

Gracias a la base de datos de modelos proteicos de *Trebouxia* sp. TR9 se obtuvieron 162 identificaciones de péptidos únicos en ambas algas de las que tres estaban codificadas por genes del genoma cloroplástico (rbcL, atpA y atpB) y el resto del genoma nuclear. En el caso

de *Trebouxia jamesii*, de los 162 modelos proteicos identificados en total para ambas algas, 81 modelos nucleares sólo se presentaron en *T. jamesii* (Tabla 21). Estos modelos proteicos presentaron 116 dominios PFAM y 21 familias CAZy; 3 modelos proteicos presentaron motivos de péptidos señal, de los cuales, los modelo g1793.t1 y g3020 podrían estar relacionados con los exopolímeros de *T. jamesii*. Se encontraron 4 modelos que tenían motivos de hélices transmembrana relacionados con el retículo endoplasmático y el cloroplasto. De todos los modelos identificados, 7 no obtuvieron ninguna anotación BLAST2GO (Descripción –NA– en Tabla 21), pero dos de estos modelos tenían dominios PFAM, el domino PF12738 (PTCB-BRCT), que es un dominio presente en proteínas implicadas en puntos de control y de reparación de ADN y el dominio PF02892 (BED-type zinc finger domain) encontrado en factores de regulación celular y en transposasas (Tabla 21). Muchas de estas identificaciones obtenidas corresponden a modelos proteicos de *Trebouxia* sp. TR9 que no se espera que sean parte de las proteínas extracelulares (e.g. se identificaron varias proteínas ribosomales, factores de transcripción, de ubiquitinización, cloroplásticas y mitocondriales o de proteosoma) y probablemente provengan de células que se rompieron durante el proceso de extracción de EPS. Sin embargo, se han identificado ocho polipéptidos probablemente relacionados con la pared celular y/o los EPS de este alga: el modelo peptídico g7420.t1 "aldo-keto reductase family 4 member c9-like"; el modelo peptídico g1177.t1 "alpha beta hidrolase" (con la presencia de la familia de transferasas de glicósidos 4 de la base de datos CAZy); el modelo peptídico g1438.t1 "apha-glucan-protein synthase" con el dominio PF03214 con un posible rol en la síntesis de polisacáridos de pared y/o EPS; el modelo peptídico g1793.t1 "barwin-like endoglucanase" que presenta el dominio PF01357 (Pollen allergen); el modelo peptídico g98.t1 "copper radical oxidase", perteneciente a la familia CAZy AA5 presente en oxidasas de galactosa; el modelo peptídico g558.t1 "epimerase, Xylose isomerase-like"; el modelo peptídico g8879.t1 "hydrolase" que contiene el dominio PF02836 (Glycosyl hydrolases) y las familia CAZy CGM42 relacionada a funciones como glycoside hydrolases y glycosyltransferases; el modelo g637.t1 "probable rhamnose biosynthetic enzime 1-like" con el dominio PF01370 y las familias GH1 y GT2 de CAZy presentes en diferentes epimerasas y por último, el modelo g8685.t1 anotado como una "tiorredoxina" que presenta el dominio PF00085 de tiorredoxinas y la familia Glycosyl Transferase Family 90 de CAZy (Tabla 21).

En el alga *Trebouxia* sp. TR9, de las 81 identificaciones obtenidas con los modelos proteicos nucleares, 33 fueron únicos para esta especie (Tabla 22). Estos modelos proteicos presentaron en total 47 dominios PFAM y 4 familias CAZy. Tan solo 3 identificaciones presentaron péptido señal y una de ellas motivos transmembrana (Tabla22). De las dos identificaciones con modelos proteicos que no obtuvieron anotación

Cuadro 21: **Identificación de las proteínas totales (digestión tríptica en líquido) presentes en los EPS de *Trebouxia jamesii*.**

| ID | Descripción | Dominios PFAM | Familias Cazy | SignalP | TMHMMg |
|---|---|---|---|---|---|
| 98.t1 | copper radical oxidase | PF09118 | AA5 | - | - |
| g178.t1 | 60s ribosomal protein l11-2-like | PF00281, PF00673 | - | - | - |
| g228.t1 | ubiquitin-nedd8-like protein rub1-like | PF00240 | AA10 | - | - |
| g278.t1 | alkaline phosphatase | PF09423 | GH18 | - | - |
| g496.t1 | ---NA--- | PF12738 | - | - | - |
| g545.t1 | 60s ribosomal protein l3-like | PF00297 | - | - | - |
| g558.t1 | epimerase | PF01261 | GH20 | - | - |
| g618.t1 | eukaryotic initiation factor 4a-2-like | PF00270, PF00271 | CE11 | - | - |
| g637.t1 | probable rhamnose biosynthetic enzyme 1-like | PF01370 | GT2, GH1 | - | - |
| g781.t1 | phospholipase sgr2-like | PF02862 | - | - | - |
| g909.t1 | 26s protease regulatory subunit s10b homolog | PF00004 | - | - | - |
| g983.t1 | enoyl-[acyl-carrier-protein] reductase | PF13561 | GH92 | - | - |
| g998.t1 | ---NA--- | - | - | + | + |
| g1000.t1 | ---NA--- | - | - | - | - |
| g1101.t1 | 6-phosphogluconate decarboxylating 3 | PF00393, PF03446, PF10184 | GT30 | - | - |
| g1177.t1 | alpha beta hydrolase | PF00561, PF12697 | GT4 | - | - |
| g1203.t1 | coiled-coil domain-containing protein 147 | - | - | - | - |
| g1299.t1 | catalase isozyme 1-like | PF00199, PF06628 | - | - | - |
| g1370.t1 | 60s ribosomal protein l18-3-like | PF00828 | - | - | - |
| g1408.t1 | fact complex subunit spt16-like | PF00557, PF08512, PF08644, PF14826 | GH84 | - | - |
| g1438.t1 | alpha- -glucan-protein synthase | PF03214 | GT75 | - | - |
| g1467.t1 | calcium-transporting endoplasmic reticulum | PF00012, PF00122, PF00689, PF00690, PF00702 | CBM13, GH65 | - | + |
| g1656.t1 | 60s ribosomal protein l35a-1 | PF01247 | - | - | - |
| g1692.t1 | 60s ribosomal protein l9-like | PF00347 | - | - | - |
| g1720.t1 | ---NA--- | - | - | - | - |
| g1793.t1 | barwin-like endoglucanase | PF01357, PF03330 | - | + | - |
| g2016.t1 | 3 -cyclic-nucleotide phosphodiesterase | PF00233 | - | - | - |
| g2162.t1 | s-adenosylmethionine synthase 1-like isoform | PF00438, PF02772, PF02773 | - | - | - |
| g2465.t1 | aspartate aminotransferase | PF00155, PF01342 | - | - | - |
| g2754.t1 | cytosolic phosphoglucose isomerase | PF00069, PF00153, PF00342 | - | - | - |
| g2858.t1 | gram domain-containing protein 1a | - | - | - | + |
| g2928.t1 | lysosomal pro-x carboxypeptidase | PF00081, PF02777, PF05577 | - | - | - |
| g2959.t1 | phosphoribulokinase family protein | PF00485, PF07807, PF07808 | - | - | - |
| g3020.t1 | desiccation-related protein pcc13-62-like | PF13668 | - | + | - |
| g3277.t1 | nuf2 subfamily protein | PF03800 | - | - | - |
| g3392.t1 | gdt1-like protein chloroplastic-like | PF01169 | - | - | + |
| g3541.t1 | malate chloroplastic-like | PF00056, PF02866 | - | - | - |
| g3558.t1 | flagellar associated protein | PF13864 | - | - | - |
| g3816.t1 | inositol-3-phosphate synthase | PF01658, PF07994 | - | - | - |
| g4150.t1 | 40s ribosomal protein s5 | PF00177 | - | - | - |
| g4312.t1 | beta tubulin | PF00091, PF03953 | - | - | - |
| g4358.t1 | 40s ribosomal protein sa-like | PF00318 | - | - | - |
| g4360.t1 | subfamily rna polymerase sigma-70 subunit | PF04539, PF04542, PF04545 | - | - | - |
| g5214.t1 | malate mitochondrial-like | PF00056, PF02866 | - | - | - |
| g5585.t1 | chaperone protein dnaj 10-like | PF14308 | - | - | - |
| g5650.t1 | 40s ribosomal protein s9-2-like | PF00163, PF01479 | - | - | - |
| g5726.t1 | hypothetical protein COCSUDRAFT_56439 | - | - | - | - |
| g5865.t1 | glutathione s-transferase | PF00043, PF02798, PF08450 | - | - | - |
| g5886.t1 | smp-30 gluconolaconase lre domain protein | - | - | - | - |
| g6014.t1 | transcription factor | - | - | - | - |
| g6016.t1 | transcription factor | - | - | - | - |
| g6158.t1 | mitochondrial outer membrane protein porin 4-like | PF01459 | - | - | - |
| g6180.t1 | glutathione chloroplastic-like | PF00070, PF02852, PF07992 | AA3 | - | - |
| g6543.t1 | ---NA--- | PF02892 | - | - | - |
| g6565.t1 | glutathione s-transferase a-like | PF00043, PF02798 | - | - | - |
| g6836.t1 | nucleoside diphosphate kinase | PF00334 | - | - | - |
| g6852.t1 | gtp-binding nuclear protein ran-3-like | PF00071 | - | - | - |
| g6877.t1 | ---NA--- | - | - | - | - |
| g6907.t1 | aspartic proteinase a1-like | PF00026, PF03489, PF05184 | - | - | - |
| g6952.t1 | dihydroorotase | PF01585, PF01979 | - | - | - |
| g7230.t1 | atp-dependent dna helicase q-like 1-like | PF00270, PF00271 | CE11 | - | - |
| g7286.t1 | 60s ribosomal protein l7a-like | PF00069, PF01248 | CBM13 | - | - |
| g7327.t1 | heat shock cognate 70 kda protein 2-like | PF00012 | CBM13 | - | - |
| g7364.t1 | carbonate dehydratase | PF00484 | - | - | - |
| g7367.t1 | nucleoside diphosphate kinase 1-like | PF00334 | - | - | - |
| g7420.t1 | aldo-keto reductase family 4 member c9-like | PF00248 | - | - | - |
| g7485.t1 | serine hydroxymethyltransferase 2 | PF00464 | - | - | - |
| g7659.t1 | cytochrome c | PF00034 | - | - | - |
| g8034.t1 | 40s ribosomal protein s13-like | PF00312, PF08069 | - | - | - |
| g8192.t1 | 60s ribosomal protein l4-like | PF00573, PF14374 | - | - | - |
| g8197.t1 | fructose- - chloroplastic-like | PF00316 | - | - | - |
| g8288.t1 | 60s ribosomal protein l13 | PF01294 | - | - | - |
| g8324.t1 | -like antibiotic response protein | PF12585 | - | - | - |
| g8401.t1 | 60s acidic ribosomal protein p0-like | PF00428, PF00466 | - | - | - |
| g8403.t1 | 40s ribosomal protein s4 | PF00900, PF01479, PF08071 | - | - | - |
| g8411.t1 | 60s ribosomal protein l34-like | PF01199 | - | - | - |
| g8685.t1 | thioredoxin | PF00085 | GT90 | - | - |
| g8792.t1 | nucleoside diphosphate kinase chloroplast | PF00334 | - | - | - |
| g8843.t1 | phosphoglycerate kinase | PF00162, PF02672 | - | - | - |
| g8874.t1 | ---NA--- | - | - | - | - |
| g8879.t1 | hydrolase | PF02836 | GH2, CBM42 | - | - |

Cuadro 22: **Identificación de las proteínas totales (digestión tríptica en líquido) presentes en los EPS de *Trebouxia* sp. TR9.**

| ID | Descripción | Dominios PFAM | Familias Cazy | SignalP | TMHMM |
|---|---|---|---|---|---|
| g120.t1 | probable histone-lysine n-methyltransferase | PF00856 | - | - | - |
| g182.t1 | probable voltage-gated potassium channel | PF00248 | - | - | - |
| g586.t1 | 60s ribosomal protein l19-1-like | PF01280 | - | - | - |
| g1420.t1 | 60s ribosomal protein l8-like | PF00181, PF03947 | - | - | - |
| g1674.t1 | 2-cys peroxiredoxin chloroplastic-like | PF00578, PF10417 | GH18 | - | - |
| g1816.t1 | abc transporter | PF12850, PF13558 | - | - | - |
| g2096.t1 | chaperonin cpn60- mitochondrial-like | PF00118 | - | - | - |
| g2168.t1 | ferritin- chloroplastic | PF00210, PF13345 | CBM50 | - | - |
| g2172.t1 | glycoside hydrolase | PF00150 | CBM2 | + | + |
| g2434.t1 | coiled-coil domain-containing protein 176 | - | - | - | - |
| g2628.t1 | superoxide dismutase | PF00081, PF02777 | - | - | - |
| g2685.t1 | hisitidine kinase | PF07080 | - | - | - |
| g2928.t1 | lysosomal pro-x carboxypeptidase | PF00081, PF02777, PF05577 | - | - | - |
| g3007.t1 | cell division cycle protein 48 homolog | PF00004, PF02359, PF02933, PF09336, PF11360 | - | - | - |
| g3954.t1 | centrosomal protein of 135 kda | - | - | - | - |
| g4065.t1 | superoxide dismutase | PF00081, PF00293, PF02777 | - | - | - |
| g4169.t1 | elongation factor 1-alpha-like | PF00009, PF03143, PF03144 | - | - | - |
| g5071.t1 | 40s ribosomal protein s14 | PF00411 | - | - | - |
| g5153.t1 | hypothetical protein COCSUDRAFT_49374 | - | - | - | - |
| g5154.t1 | 60s ribosomal protein l17-2-like | PF00237 | - | - | - |
| g5238.t1 | phosphomethylpyrimidine chloroplastic-like | PF01964, PF13667 | - | - | - |
| g5313.t1 | hypothetical protein COCSUDRAFT_63737 | - | - | - | - |
| g5565.t1 | calmodulin | PF13499 | CBM20 | - | - |
| g5717.t1 | midasin-like isoform x7 | - | - | - | - |
| g5744.t1 | acyl- oxidase | PF01756, PF02770 | - | - | - |
| g5853.t1 | ---NA--- | - | - | - | - |
| g6277.t1 | elongation factor 2-like | PF00679, PF03144, PF03764, PF14492 | - | - | - |
| g6279.t1 | hypothetical protein | - | - | + | - |
| g6886.t1 | ---NA--- | PF11034 | - | - | - |
| g7554.t1 | phosphatidylglycerol phosphatidylinositol | PF02221 | - | - | - |
| g7561.t1 | aldehyde dehydrogenase family 2 member mitochondrial | PF00171 | - | - | - |
| g7604.t1 | mismatched base pair and cruciform dna re | - | - | + | - |
| g7627.t1 | ubiquitin-conjugating enzyme e2 j2-like | PF00179, PF01451, PF13414 | - | - | - |

con BLAST2GO, una de estas, presentó un domino PFAM (PF11034, DUF2823), aunque esta familia de proteínas parecen ser reprimidas por glucosa, su función aún es desconocida. Al igual que en *T. jamesii*, la mayoría de identificaciones han sido con modelos proteicos de origen citoplasmático, cloroplástico o mitocondrial. El modelo g2172.t1 es probable que sea una proteína perteneciente a la fracción extracelular, se trata de una hidrolasa de glicósidos de la familia 5 (PF00150) que contiene la familia CBM2 de la base de datos de enzimas activas de carbohidratos (CAZy) con función de unión a celulosa, quitina o a hemicelulosas como los xilanos. Además, presenta un péptido señal N-terminal para entrar en la ruta de secreción y un motivo de hélices transmembrana, por lo que podría situarse en la cara externa de la membrana celular. La proteína g7554.t1 es posible que corresponda a otro exopolipéptido ya que contiene un péptido señal para entrar en la ruta de secreción proteica y el dominio PF02221 (MD-2-related lipid-recognition (ML)) que está implicado en el reconocimiento de lípidos, particularmente los relacionados con patógenos (Tabla 22). Finalmente, se identificaron tres modelos proteicos de *Trebouxia* sp. TR9 de función desconocida (g6279.t1, g5153.t1 y g5313.t1) de los cuales el modelo g6279.t1 contiene un péptido señal para entrar en la ruta de secreción y que tiene similitud con proteínas de membrana.

Del total de identificaciones obtenidas, 48 modelos proteicos codificados en el genoma nuclear de *Trebouxia* sp. TR9 fueron comunes para ambas algas (Tabla 23). Estos modelos proteicos presentaron 67 dominios PFAM y 12 familias de la base de datos CAZy, 2 de los modelos presentaron un péptido señal para la ruta de secreción. Se han

Cuadro 23: **Identificación de las proteínas totales (digestión tríptica en líquido) presentes en los EPS de *Trebouxia* sp. TR9 y *T. jamesii*.**

| ID | Descripción | Dominios PFAM | Familias Cazy | SignalP | TMHMM |
|---|---|---|---|---|---|
| g71.t1 | histone h2b | PF00125 | - | - | - |
| g157.t1 | ---NA--- | - | - | - | - |
| g285.t1 | probable udp-n-acetylglucosamine--peptide n-acetylglucosaminyltransferase spindly-like isoform x1 | PF00504, PF00515, PF13414, PF13844 | GT41 | - | - |
| g535.t1 | s-adenosyl homocysteine hydrolase | PF00670, PF05221 | - | - | - |
| g816.t1 | oxygen-evolving enhancer protein chloroplast | PF01716 | - | - | - |
| g831.t1 | 40s ribosomal protein s3-3-like | PF00189, PF07650 | - | - | - |
| g903.t1 | peptidyl-prolyl cis-trans isomerase | PF00160 | GT66 | - | - |
| g1382.t1 | glyceraldehyde-3-phosphate dehydrogenase | PF00044, PF02800 | - | - | - |
| g1455.t1 | enolase | PF00113, PF03952 | - | - | - |
| g1596.t1 | heat shock cognate 70 kda protein 2-like | PF00012 | CBM13 | - | - |
| g1843.t1 | 5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase | PF01717, PF08267 | - | - | - |
| g2366.t1 | flagellar flavodoxin | PF03358 | AA6 | - | - |
| g2521.t1 | porphobilinogen chloroplastic-like | PF01379, PF03900 | - | - | - |
| g2540.t1 | 40s ribosomal protein s2-4-like | PF00333, PF03719 | - | - | - |
| g2702.t1 | aldehyde dehydrogenase | PF00171 | - | - | - |
| g2758.t1 | elongation factor 1-alpha-like | PF00009, PF03143, PF03144 | - | - | - |
| g2863.t1 | 40s ribosomal protein s18-like | PF00416 | - | - | - |
| g2884.t1 | hypothetical protein COCSUDRAFT_67654 | PF03330 | - | + | - |
| g3721.t1 | ---NA--- | - | - | - | - |
| g4435.t1 | transketolase 7 | PF00456, PF02779, PF02780 | - | - | - |
| g4837.t1 | heme-binding protein 2-like | PF04832 | - | + | - |
| g4928.t1 | ribulose- -bisphosphate carboxylase oxygenase small subunit | PF00101 | - | - | - |
| g5105.t1 | 40s ribosomal protein s15-like | PF00203, PF08317 | - | - | - |
| g5227.t1 | glycoside hydrolase | PF00150, PF04012 | GH5 | - | - |
| g5385.t1 | 60s ribosomal protein l12 | PF00298, PF03946 | - | - | - |
| g5407.t1 | guanine nucleotide-binding protein subunit | PF00400 | GT77 | - | - |
| g5844.t1 | probable fructose-bisphosphate aldolase chloroplastic | PF00274 | - | - | - |
| g5982.t1 | monodehydroascorbate reductase-like | PF00070, PF04884, PF07992 | AA3 | - | - |
| g5997.t1 | hypothetical protein CHLNCDRAFT_135735 | PF00168 | - | - | - |
| g6003.t1 | hypothetical protein CHLNCDRAFT_135735 | - | - | - | - |
| g6085.t1 | triosephosphate cytosolic-like | PF00121 | - | - | - |
| g6123.t1 | phosphoglucomutase family protein | PF00408, PF02878, PF02879, PF02880 | - | - | - |
| g6347.t1 | ---NA--- | - | - | - | - |
| g7215.t1 | udp-galactopyranose mutase | PF13450 | GT4 | - | - |
| g7392.t1 | 14-3-3 protein | PF00244 | - | - | - |
| g7908.t1 | ru large subunit-binding protein subunit | PF00118 | - | - | - |
| g7949.t1 | heat shock protein 83-like | PF00183, PF13589 | GT2 | - | - |
| g8174.t1 | tubulin alpha chain | PF00091, PF03953 | - | - | - |
| g8180.t1 | alpha-l-arabinofuranosidase b | PF14587 | GH30 | - | - |
| g8214.t1 | actin actin-like protein | PF00022 | - | - | - |
| g8294.t1 | nascent polypeptide-associated complex subunit alpha-like protein 1 | PF01849 | - | - | - |
| g8398.t1 | hypothetical protein COCSUDRAFT_67654 | PF03330 | - | - | - |
| g8630.t1 | glyceraldehyde-3-phosphate dehydrogenase | PF00044, PF02800 | - | - | - |
| g8887.t1 | glutaredoxin family protein | PF00462 | - | - | - |
| g9212.t1 | -alpha-glucan-branching enzyme 2-chloroplastic amyloplastic-like | PF00128, PF00316, PF02922 | CBM48, GH13 | - | - |

identificado 3 modelos proteicos de *Trebouxia* sp. TR9 que no obtuvieron ningún tipo de anotación (–NA– en la Tabla 23). Al igual que en los casos anteriores, muchas de las identificaciones son con modelos con una posible localización en el citoplasma y orgánulos. Los modelos proteicos comunes que podrían estar relacionados con péptidos extracelulares son: el modelo g8180.t1 anotado como una "alpha-l-arabinofuranosidase b" que presenta el dominio PFAM PF14587 (glycoside hydrolase family 30) y la familia GH30 de la base de datos CAZy correspondiente a una enzima que cataliza la hidrólisis de enlaces glicosídicos cuya función podría ser la de la hidrólisis de alfa-L-arabinofuranósidos en alfa-L-arabinósidos; el modelo peptídico g6123.t1 es una posible proteína de la familia de las Phosphoglucomutase/phosphomannomutase y el modelo g7215.t1 obtenido en las identificaciones que es una posible "udp-galactopyranose mutase" y contiene el motivo PFAM PF13450 (NAD(P)-binding Rossmann-like) de unión a NAD(P) y pertenece a la familia de la base de datos CAZy de transferasas de glicósidos GT4 (Tabla 23). Finalmente, dos de las identificaciones fueron con los modelos proteicos de *Trebouxia* sp. TR9 g2284.t1 y g8398.t1 de función desconocida y similares a dos proteínas de *Coccomyxa subellipsoidea* que presentaron el dominio PF03330, este dominio está presente en proteínas de función desconocida, pero que se ha encontrado en la fracción N-terminal de alérgenos de polen (Ivanciuc *et al.* , 2009).

*Identificación de péptidos extracelulares totales de bandas de gel monodimensional*

Todas las bandas bandas recortadas del gel monodimensional de la Figura 54 han podido ser identificadas empleando la base de datos de los modelos proteicos de *Trebouxia* sp. TR9. En *Trebouxia jamesii*, al no presentarse cambios en el patrón de bandas ante ambos tratamientos, tan sólo se cortaron y secuenciaron las bandas del tratamiento control de pesos diferentes a las presentes en *Trebouxia* sp. TR9. Los resultados de las identificaciones de todas las bandas se pueden observar en la Tabla 24:

La banda 1 de *T. jamesii* , es la sub-unidad grande de la RuBisCO codificada en el genoma plastidial, este polipéptido también fue identificado en las bandas de *Trebouxia* sp. TR9 0 µM nº 5 y 100 µM nº 8. Estos polipéptidos intracelulares probablemente se originaron por la ruptura de células durante la extracción de EPS, al igual que, los polipéptidos de *T. jamesii* de las bandas nº 3, identificado como una chaperona plastidial, la nº 6, relacionado con ATPasas de segregación de cromosomas y las bandas 8 y 9 correspondientes a proteínas ribosomales. Sin embargo, las bandas restantes muy probablemente correspondan a proteínas genuinamente extracelulares. Así pues, en la banda 2 de *T. jamesii* se ha identificado la misma "glicoside hydrolase" (g2172.t1) que aparece en dos bandas de *Trebouxia* sp. TR9, la banda 2.1 en el tratamiento 0 µM y la banda 6.2 en el tratamiento 100 µM de Pb. De ellas, la banda 6.2 de *Trebouxia* sp. TR9 mantiene un tamaño similar a la banda 2 de *T. jamesii*. En este polipéptido de *Trebouxia* sp. TR9 se ha detectado la presencia de un péptido señal, de dominios transmembrana y el dominio PFAM "Cellulase" ("glycosyl hydrolase family 5"). La banda 4, que se identificó como una "Cu radical oxidase", puede estar relacionada con la generación de peróxido extracelular y presenta los dominios AA5 (Familia de actividad auxiliar de oxidasas de cobre) y DUF1929 (de función desconocida, pero encontrado en enzimas que utilizan azúcar, tales como la galactosa oxidasa) de las bases de datos CAZy y PFAM, respectivamente. La banda 5 de *T. jamesii*, al igual que la banda 6 de *Trebouxia* sp. TR9 control fueron identificadas como "glycoside hydrolase", contiene el motivo PFAM "Cellulase" y la familia CAZy "glycosyl hydrolase family 5" en su secuencia. La banda 7 se identificó con un modelo peptídico de *Trebouxia* sp. TR9 que constaba de tres dominios del PFAM: PF02922, "Carbohydrate-binding module 48" ("Isoamylase N-terminal domain"); PF00316, "Fructose-1-6-bisphosphatase" y PF00128 "Alpha amylase, catalytic domain".

La banda 1 de *Trebouxia* sp. TR9 se identificó con un modelo peptídico de *Trebouxia* sp. TR9 que contiene el motivo PFAM PF01469 de proteínas de función desconocida que contienen Pentapeptide repeats. Además TMHHM predijo un firma de unión a membrana, por lo que es posible que este péptido se encuentre unido a membranas.

La banda 2 produjo una serie de péptidos trípticos que dieron como resultado tres identificaciones diferentes: (a) 2.1 correspondiente al modelo g2172.t1 que es probablemente una celulasa también presente en la banda 2 de *T. jamesii*; (b) 2.2 identificado con un modelo de función desconocida, que contiene un péptido señal transmembrana y (c) 2.3, que presentó la mayor cobertura de las tres, correspondería a una proteína relacionada con la respuesta a la deshidratación, la cual contiene el motivo PFAM PF13668 del tipo ferritina. La banda 3 también obtuvo la misma identificación que la banda 2 de *T. jamesii*. La banda 4 fue identificada con el modelo peptídico g8398.t1 de *Trebouxia* sp. TR9, semejante a una proteína de *Coccomyxa subellipsoidea* que contiene el motivo PF03330 ("Rare lipoprotein A (RlpA)-like double-psi beta-barrel"), pero con una cobertura muy baja. La banda 5 también obtuvo la misma identificación que la banda 1 de *T. jamesii*. La banda 6 presentó también una serie cuatro identificaciones diferentes: (a) 6.1 se identificó con el modelo g2172.t1 al igual que en la banda 2 de *T. jamesii*; (b) 6.2 se identificó con el modelo g6279.t1 de función desconocida, pero que con Sinal-P se identificó un péptido señal para entrar en la ruta de transporte y/o secreción; (c) 6.3 fue identificada con una proteína asociada a flagelo que contiene el dominio conservado cd01141 ("TroA_d Periplasmic binding protein") y un péptido señal identificado por Signal-P; (d) 6.4 se trataría de una proteína semejante a las moléculas de adhesión de procariotas que pueden ocurrir en pili (fimbrias), flagelos, o en la superficie celular y que contiene una firma de proteínas embebidas en la membrana obtenida por TMHMM. La banda 7 produjo dos identificaciones 7.1 y 7.2 relacionadas con una deshidrogenasa de aldehídos. La banda 8 también obtuvo la misma identificación de la banda 1 de *T. jamesii*. La banda 9 se identificó con una carboxipeptidasa lisosomal con tres dominios PFAM: PF05577 ("Serine carboxypeptidase S28"), PF00081 ("Iron/manganese superoxide dismutases, alpha-hairpin domain") y PF02777 ("Iron/manganese superoxide dismutases, C-terminal domain").

### 4.6.2 *Discusión*

En coordinación con el grupo Plantstres de la Universidad de Alcalá de Henares, se ha podido continuar con el estudio de los exoproteomas de *Trebouxia* sp. TR9 y *T. jamesii* realizados en el trabajo de Casano et al. *(2015)*. En dicho trabajo, ambas algas fueron cultivadas durante una semana ante 0 y 100 µM $Pb(NO_3)_2$y se obtuvieron las proteínas extracelulares que, gracias a la secuenciación y anotación de los genomas de *Trebouxia* sp. TR9 realizados en la presente Tesis, pudieron ser identificados por espectrometría de masas. Para realizar la secuenciación de proteínas de *Trebouxia* sp. TR9 y *T. jamesii*, se optaron por dos aproximaciones experimentales, la secuenciación proteica de los extractos totales de proteínas extracelulares de los tratamientos control

Cuadro 24: **Identificación de péptidos extracelulares de *Trebouxia* sp. TR9 y *T. jamesii* tratadas con 0 o 100 µM $Pb(NO_3)_2$.** Cada banda fue recortada de una separación electroforética en gel de acrilamida SDS marcadas en la Figura 54. Cuando diferentes modelos proteicos de *Trebouxia* sp. TR9 se obtuvieron en la identificación de una misma banda, se indica con un segundo número.

| Organismo y tratamiento | Banda | Identificador | Descripción | Cobertura de secuencia |
|---|---|---|---|---|
| *Trebouxia jamesii* 0 µM | Banda 1 | TR9_CP_rbcL | ribulose-bisphosphate carboxylase oxygenase large subunit | 42% |
| | Banda 2 | g2172.t1 | glycoside hydrolase | 3% |
| | Banda 3 | g6118.t1 | chaperone protein chloroplastic-like | 36% |
| | Banda 4 | g98.t1 | copper radical oxidase | 13% |
| | Banda 5 | g5227.t1 | glycoside hydrolase | 16% |
| | Banda 6 | g6347.t1 | ---NA--- | 32% |
| | Banda 7 | g9212.t1 | -alpha-glucan-branching enzyme 2- chloroplastic amyloplastic-like | 22% |
| | Banda 8 | g1273.t1 | 54s ribosomal protein mitochondrial-like | 77% |
| | Banda 9 | g8294.t1 | nascent polypeptide-associated complex subunit alpha-like protein 1 | 35% |
| *Trebouxia* sp. TR9 0 µM | Banda 1 | g3233.t1 | ---NA--- | 8% |
| | Banda 2.1 | g2172.t1 | glycoside hydrolase | 5% |
| | Banda 2.2 | g6279.t1 | hypothetical protein | 10% |
| | Banda 2.3 | g8611.t1 | desiccation-related protein pcc13-62-like | 17% |
| | Banda 3 | g5227.t1 | glycoside hydrolase | 25% |
| | Banda 4 | g8398.t1 | hypothetical protein COCSUDRAFT_67654 | 3% |
| | Banda 5 | TR9_CP_rbcL | ribulose-bisphosphate carboxylase oxygenase large subunit | 49% |
| *Trebouxia* sp. TR9 100 µM | Banda 6.1 | g3547.t1 | purple acid phosphatase 18-like | 10% |
| | Banda 6.2 | g2172.t1 | glycoside hydrolase | 5% |
| | Banda 6.3 | g7973.t1 | flagellar associated protein | 16% |
| | Banda 6.4 | g8420.t1 | adhesin-like protein | 3% |
| | Banda 7.1 | g7561.t1 | aldehyde dehydrogenase family 2 member mitochondrial-like | 21% |
| | Banda 7.2 | g2702.t1 | aldehyde dehydrogenase | 50% |
| | Banda 8 | TR9_CP_rbcL | ribulose-bisphosphate carboxylase oxygenase large subunit | 49% |
| | Banda 9 | g2928.t1 | lysosomal pro-x carboxypeptidase | 10% |

de ambas algas en digestión líquida y la secuenciación proteica de diferentes bandas obtenidas en un gel monodimensional SDS-PAGE que eran o propias de *T. jamesii*, o que su concentración ante el tratamiento con 100 µM de Pb era diferente a la del control en *Trebouxia* sp. TR9.

### *Identificación de péptidos extracelulares obtenidos por digestión en líquido*

La anotación de los genomas de *Trebouxia* sp. TR9 ha mejorado tanto cualitativa como cuantitativamente la identificación de los extractos totales de los tratamientos control de exoproteínas de *Trebouxia* sp. TR9 y *T. jamesii* en comparación con la realizada con la base de datos del NCBI. Esto es debido a que la herramienta MASCOT utiliza los valores de masa experimentales obtenidos por espectrometría de masas y los compara con los valores medios de masas de las entradas presentes en una base de datos. Si el péptido comparado es parte de una proteína presente en la base de datos de secuencias peptídicas, obtendrá esa entrada precisa como identificación. Si la base de datos de secuencias no contiene la proteína a identificar o no existen proteínas con secuencias equivalentes de especies relacionadas, no obtendrá ninguna identificación. Es por ello, que al comparar las identificaciones obtenidas frente a la base de datos del NCBI-Viridiplantae, se obtuvieron peores resultados que contra la base de datos de los modelos proteicos de *Trebouxia* sp. TR9, tanto para este alga, como para *T. jamesii.*

Al comparar el número de las identificaciones entre ambas algas, el extracto de exo-proteínas de *T. jamesii* ha obtenido un mayor número de éstas en comparación con las identificaciones obtenidas en *Trebouxia* sp. TR9, esto concuerda con los datos aportados por el trabajo de Casano *et al.* (2015), donde *T. jamesii* presentaba casi un 50 % de contenido proteico en los extractos de polímeros extracelulares frente a casi un 30 % en *Trebouxia* sp. TR9. En cada alga han aparecido componentes citoplasmáticos u organulares (Tablas 21, 22 y 23), muchos de ellos relacionados con componentes de rutas para evitar el estrés oxidativo (chaperonas, Heat Shock Proteins, catalasa, superóxido dismutasa y relacionados con el glutation), con la biosíntesis y control de proteínas (proteínas ribosomales, factores de elongación, de ubiquitinización o del proteosoma) o pertenecientes y/o codificadas en los genomas de los orgánulos (*rbcL*, *atpA* y a*tpB*). Estos resultados indican que en el proceso de extracción de las proteínas, se arrastraron componentes citoplasmáticos y organulares de las células. Además, el mayor número de polipéptidos intracelulares encontrados en *T. jamesii* podría ser debido a que su pared celular es de menor grosor u¡y posiblemente más frágil (Casano *et al.* , 2011, 2015), por lo que podría haber colapsado en muchas más ocasiones que en *Trebouxia* sp. TR9 y por tanto se arrastraron más componentes citoplasmáticos. Aun así, en las identificaciones de *T. jamesii*, se han encontrado ocho polipéptidos que posiblemente estén relacionados con la fracción exopolimérica de este microalga y que no aparecieron en *Trebouxia* sp. TR9. Recíprocamente en *Trebouxia* sp. TR9 tan solo dos exopéptidos parecen ser exclusivos de este alga ya que no están presentes en *T. jamesii*. Ambas algas presentaron 48 polipéptidos comunes, de los cuales 3 parecen estar relacionadas con el metabolismo de pared celular y dos, con modelos de función desconocida, pero parecidas a proteínas hipotéticas de *Coccomyxa subellipsoidea* que presentan un dominio del PFAM relacionado con alérgenos de polen. König & Peveling (1984) caracterizaron la pared celular de *Trebouxia* y encontraron una capa de esporopolenina, estas proteínas podrían ser las constituyentes de dicha capa.

Del total de péptidos identificados con la base de datos de modelos proteicos de *Trebouxia* sp. TR9, se han deducido 16 polipéptidos/proteínas de *Trebouxia* sp. TR9 que no tienen similitud con las secuencias depositadas hasta la fecha en el GenBank. Por este motivo en la anotación realizada basada en comparaciones de similitud con la herramienta BLAST2GO, no se encontró ningún tipo de coincidencia. Por tanto, para estos polipéptidos de *Trebouxia* sp. TR9 cuya posible existencia se ha podido corroborar experimentalmente, harán falta nuevos experimentos proteómicos, transcriptómicos y bioquímicos que permitan caracterizar y corroborar apropiadamente su presencia y función.

*Identificación de péptidos extracelulares de bandas recortadas del gel monodimensional*

En el trabajo de Casano *et al.* (2015), se extrajeron los péptidos extracelulares de *Trebouxia* sp. TR9 y *T. jamesii* y fueron separados en electroforesis monodimensional (SDS-PAGE), en dicho trabajo, se observó que ambas algas presentaban diferentes patrones polipéptidos y que ante el tratamiento de 100 µM de $Pb(NO_3)_2$ , dicho patrón cambiaba en *Trebouxia* sp. TR9, mientras que en *T. jamesii* no. Como continuación de dicho trabajo, se recortaron diferentes bandas para cada alga y con la ayuda de la base de datos de los modelos proteicos de *Trebouxia* sp. TR9 se pudieron identificar todas las bandas obtenidas. Varias de ellas fueron identificadas como la sub-unidad grande de la proteína cloroplástica RuBisCO (Banda 1 para *T. jamesi*i y las bandas 5 y 8 para *Trebouxia* sp. TR9 de la Figura 54 anotadas en la Tabla 24). Esta proteína cloroplástica, que supone cerca del 20 % de las proteínas totales de *Trebouxia* sp. TR9 (Leonardo Casano, comunicación personal), ha de ser una contaminación que se ha arrastrado durante la extracción de los exopéptidos. Las proteínas identificadas presentaron diferentes tamaños en ambas algas: en *T. jamesii* la banda 1 se encuentra cerca de los 150 kDa, mientras que las bandas identificadas en *Trebouxia* sp. TR9 se encuentran alrededor de los 30 kDa (Figura 54). Dado que la sub-unidad grande de la RuBisCO posee un tamaño de alrededor de 55 kDa en las plantas en general, en el caso de *T. jamesii*, la proteína identificada como tal, podría ser un dímero la misma y en el caso de *Trebouxia* sp. TR9, al tratarse de péptidos de un menor tamaño, es probable que sean productos de degradación. Aún así, en *Trebouxia* sp. TR9, después de tratamientos con 100 µM de Pb, aumenta la cantidad de ambas bandas (Figura 54), por lo que es posible que el fragmento de la sub-unidad grande de la RuBisCO esté enmascarando uno o varios péptidos cuya cantidad aumenta después de los tratamientos con plomo. Puesto que la separación electroforética se realizó en una sola dimensión, la presencia de SDS y la etapa de desnaturalización hacen que las proteínas se separen por su tamaño, pero puede ocurrir la migración de varios péptidos de tamaños equivalentes en una misma banda. Esto parece suceder en las bandas 2 ,6 y 7 de *Trebouxia* sp. TR9, donde se identificaron varios péptidos pertenecientes a diferentes modelos proteicos del *Trebouxia* sp. TR9 (Tabla 24). Para poder solventar este problema, se deberán realizar futuros experimentos de separación electroforética bidimensional de proteínas y su posterior secuenciación.

La banda 2 de *T. jamesii* y las bandas 2.1 y 6.2 de *Trebouxia* sp. TR9 se han identificado con el modelo peptídico g2172.t1 anotado como un hidrolasa de glicósidos (aunque la cobertura de las identificaciones fue baja, entre 3 y 5 %). La banda 2 de *T. jamesii* y la banda 6.2 de *Trebouxia* sp. TR9 presentan una migración similar en el gel, mientras que la banda 2 de *Trebouxia* sp. TR9 su tamaño es menor. Es posible

que estas bandas sean dos proteínas diferentes con un mismo motivo proteico relacionado con la hidrólisis de uniones glicosídicas de azúcares complejos y que en el caso de *Trebouxia* sp. TR9, estas proteínas aumenten su cantidad después de los tratamientos con plomo. En *T. jamesii* y en *Trebouxia* sp. TR9 las bandas 5 y 3 han sido identificadas como otra hidrolasa de glicósidos diferente a la anterior (modelo g5227.t1), respectivamente. En este caso la cobertura fue mayor que en el anterior resultado (16 % para *T. jamesii* y 25 % para *Trebouxia* sp. TR9). El tamaño de la banda de *Trebouxia* sp. TR9 es mayor y su cantidad también aumenta después de los tratamientos con plomo. Estas identificaciones son un indicio de que las EPS incluyen estas proteínas relacionadas con enzimas modificantes de polisacáridos encargadas de hidrolizar glucósidos que forman la pared. Además, en *Trebouxia* sp. TR9 la concentración de estas enzimas aumenta después de los tratamientos con plomo. Es posible que estas proteínas, forman parte de uno de los mecanismos que producen cambios en la pared celular y/o en las EPS que permiten retener extracelularmente el Pb, evitando su entrada al interior celular. En *Trebouxia* sp. TR9, la banda 7 fue anotada como dos deshidrogenasas de aldehídos que pueden estar relacionadas con procesos de detoxificación ante la presencia de Pb puesto que su concentración aumenta en *Trebouxia* sp. TR9 después de los tratamientos con sales de este metal.

En el presente estudio, se muestra por primera vez un análisis comparativo del contenido diferente de exopolipéptidos entre las algas *Trebouxia* sp. TR9 y *T. jamesii* que siempre están presentes en todos los talos de *Ramalina farinacea*. Los resultados obtenidos de la secuenciación y anotación del genoma de *Trebouxia* sp. TR9 ayudarán en futuros trabajos de secuenciación de proteínas aisladas de diferentes especies del género *Trebouxia*, ya que se han conseguido identificar cerca del doble de péptidos en comparación a los identificados utilizando la base de datos del NCBI para ambas especies. Todas las identificaciones peptídicas realizadas abren la puerta para futuros estudios sobre la tolerancia de estas algas a otros metales pesados y su implicación en la capacidad total del holobionte para sobrevivir en dichas condiciones adversas.

En una etapa posterior que excede a los objetivos de la presente Tesis, se continuará investigando la base genética de la gran plasticidad metabólica de *Trebouxia* sp. TR9. Para ello nos proponemos la realización de análisis transcriptómicos y proteómicos aplicando diversos tratamientos que reproduzcan condiciones ambientales en las que podemos encontrar el liquen *Ramalina farinacea* para lo que utilizaremos los datos genómicos aportados en la presente Tesis Doctoral.

# 5

# CONCLUSIONES FINALES

En la presente tesis doctoral, se han secuenciado y anotado los genomas mitocondrial, cloroplástico y nuclear del alga *Trebouxia* sp. TR9 mediante las tecnologías de secuenciación masiva ROCHE 454 GS FLX Titanium, ROCHE 454 GS JUNIOR paired-end e Illumina Miseq paired-end.

## 5.1 EL GENOMA MITOCONDRIAL DE *trebouxia* SP. TR9

1. Es de naturaleza circular y tiene un tamaño de 70.070 nt.

2. Se han identificado un total de 61 genes cuyo código genético fue identificado como el estándar.

3. En algunos de los genes anotados se han localizado un 9 intrones de tipo I, algunos de los cuales contienen ORFs que codifican Homing endonucleases de la familia LAGLIDADG.

4. Se ha observado un bajo nivel de sintenia en las algas de la clase Trebouxiophyceae que indicaría que se han podido producir intensos reordenamientos a lo largo de la evolución de estos organismos.

5. Además se ha comparado la presencia de genes de los genomas mitocondriales de algas verdes, siendo los genomas mitocondriales de las algas de la Clase Trebouxiophyceae con las que el contenido en genes se asemeja más al de *Trebouxia* sp. TR9, a excepción del alga parásita del género *Helicosporidium*.

6. El análisis filogenético basado en las secuencias de las proteínas codificadas por siete genes del genoma mitocondrial de *Trebouxia* sp. TR9 y los de 25 especies de algas verdes, coincide con la realizada en base a genes cloroplásticos de Lemieux *et al.* (2014) por lo que se refiere a la posición relativa de las clases Prasinophyceae, Chlorophyceae, Trebouxiophyceae y Ulvophyceae. Sin embargo, han aparecido dos diferencias fundamentales: (i) la monofilia de la clase Trebouxiophyceae y (ii) la posición de la clase Pedinophyceae más relacionada con las Chlorophyceae y Ulvophyceae.

## 5.2 EL GENOMA CLOROPLÁSTICO DE *trebouxia* SP. TR9

1. Es de naturaleza circular y que posee la estructura cuatripartita típica de los cloroplastos de plantas terrestres, en las regiones repetidas invertidas o IR incluyen un único gen, el rbcL.

2. El tamaño final del genoma cloroplástico obtenido ha sido de 303.323 nt siendo uno de los mayores tamaños conocidos en el contexto de las algas verdes de la división Chlorophyta.

3. Se han identificado un total de 108 genes, tres de los cuales codifican posibles ORFs que no presentan similitud con ninguna proteína de la base de datos del NCBI.

4. Se han identificado un total de 12 intrones de tipo I presentes en algunos genes, que es una cantidad elevada en las algas verdes. En algunos de los intrones secuenciados se han identificado 6 ORFs que podrían codificar Homing endonucleasas de la familia LAGLIDADG.

5. La comparación de los genes presentes en el genoma plastidial de *Trebouxia* sp. TR9 con los de 34 especies de algas de la clase Trebouxiophyceae, ha mostrado que los genes ycf47 e ycf62 están ausentes en 5 especies, mientras que el gen ycf20 está presente en todas las especies analizadas. Sin embargo, en otras algas de otras clases de la división Chlorophyta, la presencia de estos genes en el genoma cloroplástico poco frecuente.

## 5.3 EL GENOMA NUCLEAR DE *trebouxia* SP. TR9

1. Su tamaño se ha estimado en 40-50 Mb en base a dos aproximaciones experimentales: (i) la técnica de PCR en Tiempo Real que indica que el número de copias del gen del ARN ribosomal se aproxima a 4 (3,91 ± 0,51); (ii) el ensamblaje de los "contigs" obtenidos por secuenciación masiva que dio como resultado el recuento de un total de 59.121.427 nt. El ensamblaje más óptimo del genoma nuclear abarcaba 2.626 "contigs" que tenían una longitud N50 y N95 de 142.866 nt y 21.727 nt, respectivamente.

2. Para comprobar la continuidad del ensamblaje nuclear se ha utilizado el espacio génico calculado con la herramienta CEGMA. El genoma nuclear comprendía un 91 y 97 % del conjunto de 248 CEGs de forma completa y parcial, respectivamente.

3. Se han encontrado elementos móviles, la mayoría de ellos son retrotransposones con terminaciones terminales largas (LTR), de los cuales, los más abundantes son de tipo Gypsy/DIRS1 y Ty1/Copia.

4. La predicción de genes “ab initio” del genoma nuclear realizada con el programa AUGUSTUS, entrenado con los modelos obtenidos con el programa CEGMA, dio como resultado 9.499 posibles modelos génicos, 6.364 fueron anotados con al menos un término de Gene Ontology (GO) y de ellos, 2.249 presentaron un número enzimático asociado.

5. La clasificación del enriquecimiento de términos GO mostró que las funciones moleculares “transferase activity” e “hydrolase activity”, los componentes celulares “protein complex”, “membrane-

enclosed lumen" y "nucleus" y los procesos biológicos "biosynthetic process", "nitrogen compound metabolic process", "small molecule metabolic process", "cellular macromolecule metabolic process", "gene expression", "catabolic process", "protein metabolic process" y "macromolecule modification" son los más enriquecidos.

6. Se han obtenido un total de 10.922 dominios proteicos PFAM presentes en 6.544 modelos proteicos de los que 2.147 y 2.979 son compartidos por todas o con al menos una de las especies de algas analizadas (*Asterochloris* sp., *Chlorella variabilis*, *Coccomyxa subellipsoidea* y *Chlamydomonas reindhartii)*, respectivamente.

7. Se han encontrado 19 motivos PFAM específicos de la división Chlorophyta, 6 motivos propios de las algas liquénicas *Trebouxia* sp. TR9 y *Asterochloris* sp y 23 motivos propios de *Trebouxia* sp. TR9.

8. La identificación de proteínas relacionadas con la asimilación de carbono sugiere que *Trebouxia* sp. TR9 puede tener mecanismos de concentración de carbono de tipo C3 y C4/CAM junto con otros de fotorrespiración.

9. Se han estudiado las enzimas relacionadas con el metabolismo de carbohidratos presentes en los modelos proteicos de *Trebouxia* sp. TR9 y otras algas de la clase Trebouxiophyceae mediante dos algoritmos diferentes: (i) búsquedas de secuencias ortólogas y (ii) por la presencia de motivos PFAM y las comunes a ambos tipos de búsqueda.

10. El estudio de enzimas relacionadas con el metabolismo de carbohidratos presentes en los modelos proteicos de *Trebouxia* sp. TR9 y otras algas de la clase Trebouxiophyceae indica que entre el 11 y el 12 % de las proteínas totales de estas algas pertenecían al menos a un clan de las familias de la base de datos "Carbohydrate-Active enZYmes Database" (CAZy).

11. La anotación del genoma nuclear ha sido utilizada para apoyar el análisis proteómico de exoproteínas (EPS) realizado por Casano *et al.* (2015) de las algas liquénicas *Trebouxia* sp. TR9 y *T. jamesii*, mejorando la identificación de péptidos realizada utilizando la base de datos del NCBI.

# 6

# BIBLIOGRAFÍA

# BIBLIOGRAFÍA


Abascal, F., Zardoya, R., & Posada, D. 2006. GenDecoder: genetic code prediction for metazoan mitochondria. *Nucleic acids research*, **34**(suppl 2), W389–W393.

Adams, K.L., Daley, D.O., Whelan, J., & Palmer, J.D. 2002. Genes for two mitochondrial ribosomal proteins in flowering plants are derived from their chloroplast or cytosolic counterparts. *The Plant Cell Online*, **14**(4), 931–943.

Ahmadjian, V. 1993. *The lichen symbiosis.* New York, New York, USA: John Wiley & Sons, Inc.

Ahmadjian, V., & Jacobs, J.B. 1981. Relationship between fungus and alga in the lichen *Cladonia cristatella* Tuck. *Nature*, **289**, 169 – 172.

Alscher, R.G., Erturk, N., & Heath, L.S. 2002. Role of superoxide dismutases (SODs) in controlling oxidative stress in plants. *Journal of experimental botany*, **53**(372), 1331–1341.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, **25**(17), 3389–3402.

Álvarez, R., del Hoyo, A., García-Breijo, F., Reig-Armiñana, J., del Campo, E.M., Guéra, A., Barreno, E., & Casano, L.M. 2012. Different strategies to achieve Pb-tolerance by the two *Trebouxia* algae coexisting in the lichen *Ramalina farinacea*. *Journal of plant physiology*, **169**(18), 1797–1806.

Álvarez, R., Del Hoyo, A., Díaz-Rodríguez, C., Coello, A.J., Del Campo, E.M., Barreno, E., Catalá, M., & Casano, L.M. 2014. Lichen Rehydration in Heavy Metal-Polluted Environments: Pb modulates the oxidative response of both *Ramalina farinacea* thalli and Its isolated microalgae. *Microbial ecology*, **69**(3), 1–12.

Archibald, J.M. 2011. Origin of eukaryotic cells: 40 years on. *Symbiosis*, **54**(2), 69–86.

Armaleo, D., & May, S. 2009. Sizing the fungal and algal genomes of the lichen *Cladonia grayi* through quantitative PCR. *Symbiosis*, **49**(1), 43–51.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., & otros. 2004. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science*, **306**(5693), 79–86.

Asada, K. 2006. Production and scavenging of reactive oxygen species in chloroplasts and their functions. *Plant physiology*, **141**(2), 391–396.

Ausubel, F.M., Brent, R., Kingston, RE, Moore, DD, Seidman, JG, Smith, JA, & Struhl, K. 1989. *Current protocols in molecular biology*. Vol. 1. New York, New York, USA: John Wiley & Sons, Inc.

Badger, M.R., Pfanz, H., B
üdel, B., Heber, U., & Lange, O.L. 1993. Evidence for the functioning of photosynthetic CO 2-concentrating mechanisms in lichens containing green algal and cyanobacterial photobionts. *Planta*, **191**(1), 57–70.

Bateman, A., Coin, L., Durbin, ., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.LL., & otros. 2004. The Pfam protein families database. *Nucleic acids research*, **32**(suppl 1), D138–D141.

Bélanger, A.S., Brouard, J.S., Charlebois, P., Otis, C., Lemieux, C., & Turmel, M. 2006. Distinctive architecture of the chloroplast genome in the chlorophycean green alga *Stigeoclonium helveticum*. *Molecular Genetics and Genomics*, **276**(5), 464–477.

Bhattacharya, D., Qiu, H., Price, D.C., & Yoon, H.S. 2015. Why we need more algal genomes. *Journal of Phycology*, **51**, 1–5.

Blanc, G., Duncan, G., Agarkova, I., Borodovsky, M., Gurnon, J., Kuo, A., Lindquist, E., Lucas, S., Pangilinan, J., Polle, J., & otros. 2010. The *Chlorella variabilis* NC64A genome reveals adaptation to fhotosymbiosis, Coevolution with viruses, and cryptic sex. *The Plant Cell Online*, **22**(9), 2943–2955.

Blanc, G., Agarkova, I., Grimwood, J., Kuo, A., Brueggeman, A., Dunigan, D.D., Gurnon, J., Ladunga, I., Lindquist, E., Lucas, S., & otros. 2012. The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biology*, **13**(5), R39.

Bradford, M.M. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical biochemistry*, **72**(1), 248–254.

Brouard, J.S., Otis, C., Lemieux, C., & Turmel, M. 2010. The exceptionally large chloroplast genome of the green alga *Floydiella terrestris* illuminates the evolutionary history of the Chlorophyceae. *Genome Biology and Evolution*, **2**, 240–256.

Calatayud, A., Guera, A., Fos, S., & Barreno, E. 2001. A new method to isolate lichen algae by using Percoll® gradient centrifugation. *The Lichenologist*, **33**(4), 361–366.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., & Henrissat, B. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic acids research*, **37**(suppl 1), D233–D238.

Casano, L.M., Del Campo, E.M., García-Breijo, F.J., Reig-Armiñana, J., Gasulla, F., Del Hoyo, A., Guéra, A., & Barreno, E. 2011. Two *Trebouxia* algae with different physiological performances are ever-present in lichen thalli of *Ramalina farinacea*. Coexistence versus Competition? *Environmental Microbiology*, **13**(3), 806–818.

Casano, L.M., Braga, M.R., Álvarez, R., del Campo, E.M., & Barreno, E. 2015. Differences in the cell walls and extracellular polymers of the two *Trebouxia* microalgae coexisting in the lichen *Ramalina farinacea* are consistent with their distinct capacity to immobilize extracellular Pb. *Plant Science*, **236**, 195–204.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution*, **17**(4), 540–552.

Catalá, M., Gasulla, F., Pradas del Real, A.E., García-Breijo, F., Reig-Armiñana, J., & Barreno, E. 2010. Fungal-associated NO is involved in the regulation of oxidative stress during rehydration in lichen symbiosis. *BMC microbiology*, **10**(1), 297.

Catalá, M., Gasulla, F., Pradas del Real, A.E., García-Breijo, F., Reig-Armiñana, J., & Barreno, E. 2013. The organic air pollutant cumene hydroperoxide interferes with NO antioxidant role in rehydrating lichen. *Environmental Pollution*, **179**, 277–284.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., MÃŒller, W.E.G., Wetter, T., & Suhai, S. 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome research*, **14**(6), 1147.

Chou, H.H., & Holmes, M.H. 2001. DNA sequence quality trimming and vector removal. *Bioinformatics*, **17**(12), 1093–1104.

Chu, H., Jo, Y., & Cho, W.K. 2014. Evolution of endogenous non-retroviral genes integrated into plant genomes. *Current Plant Biology*, **1**, 55–59.

Conant, G.C., & Wolfe, K.H. 2008. GenomeVx: simple web-based creation of editable circular chromosome maps. *Bioinformatics*, **24**(6), 861–862.

Conesa, A., & Götz, S. 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International journal of plant genomics*, **21**(18), 3674–3676.

Cowan, I.R., Lange, O.L., & Green, T.G.A. 1992. Carbon-dioxide exchange in lichens: determination of transport and carboxylation characteristics. *Planta*, **187**(2), 282–294.

Darling, A.C.E., Mau, B., Blattner, F.R., & Perna, N.T. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, **14**(7), 1394–1403.

Darriba, D., Taboada, G.L., Doallo, R., & Posada, D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, **27**(8), 1164–1165.

De Cambiaire, J.C., Otis, C., Turmel, M., & Lemieux, C. 2007. The chloroplast genome sequence of the green alga *Leptosira terrestris*: multiple losses of the inverted repeat and extensive genome rearrangements within the Trebouxiophyceae. *BMC genomics*, **8**(1), 213.

De Clerck, O., Bogaert, K.A., & Leliaert, F. 2012. Diversity and evolution of algae: primary endosymbiosis. *Advances in Botanical Research*, **64**, 55–86.

De Koning, A.P., & Keeling, P.J. 2006. The complete plastid genome sequence of the parasitic green alga *Helicosporidium* sp. is highly reduced and structured. *BMC biology*, **4**(12).

De La Torre, R., Sancho, L.G., Horneck, G., de los Ríos, A., Wierzchos, J., Olsson-Francis, K., Cockell, C.S., Rettberg, P., Berger, T., de Vera, J.P., & otros. 2010. Survival of lichens and bacteria exposed to outer space conditions–Results of the *Lithopanspermiai* experiments. *Icarus*, **208**(2), 735–748.

del Campo, E.M., Casano, L.M., Vidal, G., & Barreno, E. 2009. Presence of multiple group I introns closely related to bacteria and fungi in plastid 23S rRNAs of lichen-forming *Trebouxia*. *International Microbiology*, **12**, 59–67.

del Campo, E.M., Hoyo, A., Casano, L.M., Martinez-Alberola, F., & Barreno, E. 2010a. A rapid and cost-efficient DMSO-based method for isolating DNA from cultured lichen photobionts. *Taxon*, **59**(2), 588–591.

del Campo, E.M., Casano, L.M., Gasulla, F., & Barreno, E. 2010b. Suitability of chloroplast LSU rDNA and its diverse group I introns for species recognition and phylogenetic analyses of lichen-forming *Trebouxia* algae. *Molecular phylogenetics and evolution*, **54**(2), 437–444.

del Campo, E.M., Catalá, S., Gimeno, J., Hoyo, A., Martínez-Alberola, F., Casano, L.M., Grube, M., & Barreno, E. 2013. The genetic structure of the cosmopolitan three-partner lichen *Ramalina farinacea* evidences the concerted diversification of symbionts. *FEMS Microbiology Ecology*, **83**(2), 310–323.

del Hoyo, A., Álvarez, R., del Campo, E.M., Gasulla, F., Barreno, E., & Casano, L.M. 2011. Oxidative stress induces distinct physiological responses in the two *Trebouxia* phycobionts of the lichen *Ramalina farinacea*. *Annals of Botany*, **107**(1), 109.

DePriest, P.T. 2004. Early molecular investigations of lichen-forming symbionts: 1986–2001*. *Microbiology*, **58**(1), 273–301.

Derelle, E., Ferraz, C., Rombauts, S., Rouzé, P., Worden, A.Z., Robbens, S., Partensky, F., Degroeve, S., Echeynié, S., Cooke, R., & otros. 2006. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proceedings of the National Academy of Sciences*, **103**(31), 11647.

Dutilh, B.E., Jurgelenaite, R., Szklarczyk, R., van Hijum, S.A.F.T., Harhangi, H.R., Schmid, M., de Wild, B.and Stunnenberg, H.G., Strous, M., Jetten, M.S.M., & otros. 2011. FACIL: fast and accurate genetic code inference and logo. *Bioinformatics*, **27**(14), 1929–1933.

Earl, D., Bradnam, K., John, J.S., Darling, A., Lin, D., Fass, J., Yu, H.O.K., Buffalo, V., Zerbino, D.R., Diekhans, M., & otros. 2011. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome research*, **21**(12), 2224–2241.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, **32**(5), 1792–1797.

Friedl, T. 1989. Comparative ultrastructure of pyrenoids in *Trebouxia* (Microthamniales, Chlorophyta). *Plant Systematics and Evolution*, **164**(1-4), 145–159.

Friedl, T., & BÃŒdel, B. 2008. Photobionts. *Chap. 2, pages 8–23 of:* Nash III, T.H. (ed), *Lichen Biology*, 2 edn. Cambridge UK: Cambridge University Press.

Friedl, T., Besendahl, A., Pfeiffer, P., & Bhattacharya, D. 2000. The distribution of group I introns in lichen algae suggests that lichenization facilitates intron lateral transfer. *Molecular phylogenetics and evolution*, **14**(3), 342–352.

Fučíková, K., Lewis, P.O., González-Halphen, D., & Lewis, L.A. 2014. Gene Arrangement Convergence, Diverse Intron Content, and Genetic Code Modifications in Mitochondrial Genomes of Sphaeropleales (Chlorophyta). *Genome biology and evolution*, **6**(8), 2170–2180.

Gasulla, F., de Nova, P., Esteban-Carrasco, A., Zapata, J.M., Barreno, E., & Guéra, A. 2009. Dehydration rate and time of desiccation affect recovery of the lichenic algae *Trebouxia erici*: alternative and classical protective mechanisms. *Planta*, **231**(1), 195–208.

Gasulla, F., Guéra, A., & Barreno, E. 2010. A simple and rapid method for isolating lichen photobionts. *Symbiosis*, **51**, 175–179.

Green, T.G.A., Nash III, T.H., & Lange, O.L. 2008. Physiological ecology of carbon dioxide exchange. *Chap. 9, pages 152–181 of:* Nash III, T.H. (ed), *Lichen Biology*, 2 edn. Cambridge UK: Cambridge University Press.

Grube, M., & Muggia, L. 2010. *Identifying algal symbionts in lichen symbioses*. Tools for Identifying Biodiversity: Progress and Problems. EUT Edizioni Università di Trieste.

Guindon, S., & Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, **52**(5), 696–704.

Hall, T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Pages 95–98 of: Nucleic acids symposium series*, vol. 41.

Halliwell, B. 2006. Reactive species and antioxidants. Redox biology is a fundamental theme of aerobic life. *Plant physiology*, **141**(2), 312–322.

Hawksworth, DL. 1988. The variety of fungal-algal symbioses, their evolutionary significance, and the nature of lichens. *Botanical Journal of the Linnean Society*, **96**(1), 3–20.

Heber, U., Lange, O.L., & Shuvalov, V.A. 2006. Conservation and dissipation of light energy as complementary processes: homoiohydric and poikilohydric autotrophs. *Journal of Experimental Botany*, **57**(6), 1211–1223.

Helms, G., Friedl, T., Rambold, G., & Mayrhofer, H. 2001. Identification of photobionts from the lichen family *Physiaceae* using algal-specific ITS rDNA sequencing. *Lichenologist*, **33**(1), 73–86.

Huerta-Cepas, J., Dopazo, J., & Gabaldón, T. 2010. ETE: a python Environment for Tree Exploration. *BMC bioinformatics*, **11**(1), 24.

Iorizzo, M., Senalik, D., Szklarczyk, M., Grzebelus, D., Spooner, D., & Simon, P. 2012. De novo assembly of the carrot mitochondrial genome using next generation sequencing of whole genomic DNA provides first evidence of DNA transfer into an angiosperm plastid genome. *BMC plant biology*, **12**(1), 61.

Ivanciuc, O., Garcia, T., Torres, M., Schein, C.H., & Braun, W. 2009. Characteristic motifs for families of allergenic proteins. *Molecular immunology*, **46**(4), 559–568.

Kappen, L. 1994. The lichen, a mutualistic system - some mainly ecophysiological aspects. *Pages 193–202 of: Crypt. Bot.*, vol. 4.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., & otros. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**(12), 1647–1649.

Keeling, P.J. 2013. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annual review of plant biology*, **64**, 583–607.

Kerney, R., Kim, E., Hangarter, R.P., Heiss, A.A., Bishop, C.D., & Hall, B.K. 2011. Intracellular invasion of green algae in a salamander host. *Proceedings of the National Academy of Sciences*, **108**(16), 6497.

Kibbe, W.A. 2007. OligoCalc: an online oligonucleotide properties calculator. *Nucleic acids research*, **35**(suppl 2), W43–W46.

Kim, K.M., Park, J., Bhattacharya, D., & Yoon, H.. 2014. Applications of next-generation sequencing to unravelling the evolutionary history of algae. *International journal of systematic and evolutionary microbiology*, **64**(Pt 2), 333–345.

Kleine, T., Maier, U. G, & Leister, D. 2009. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annual review of plant biology*, **60**, 115–138.

König, J., & Peveling, E. 1984. Cell walls of the phycobionts *Trebouxia* and *Pseudotrebouxia*: constituents and their localization. *The Lichenologist*, **16**(02), 129–144.

Kroken, S., & Taylor, J.W. 2000. Phylogenetic species, reproductive mode, and specificity of the green alga *Trebouxia* forming lichens with the fungal genus *Letharia*. *The Bryologist*, **103**(4), 645–660.

Kück, U., Jekosch, K., & Holzamer, P. 2000. DNA sequence analysis of the complete mitochondrial genome of the green alga *Scenedesmus obliquus*: evidence for UAG being a leucine and UCA being a nonsense codon. *Gene*, **253**(1), 13–18.

Lang, B.F., & Nedelcu, A.M. 2012. Plastid genomes of algae. *Pages 59–87 of:* Bock, R., & Knoop, V. (eds), *Genomics of Chloroplasts and Mitochondria*. Advances in Photosynthesis and Respiration, vol. 35. Dordrecht, Netherlands: Springer.

Le, S.Q., & Gascuel, O. 2008. An improved general amino acid replacement matrix. *Molecular biology and evolution*, **25**(7), 1307–1320.

Leliaert, F., Smith, D.R., Moreau, H., Herron, M.D., Verbruggen, H., Delwiche, C.F., & De Clerck, O. 2012. Phylogeny and molecular evolution of the green algae. *Critical Reviews in Plant Sciences*, **31**(1), 1–46.

Lemieux, C., Otis, C., & Turmel, M. 2014. Chloroplast phylogenomic analysis resolves deep-level relationships within the green algal class Trebouxiophyceae. *BMC evolutionary biology*, **14**(1), 211.

Lemieux, Claude, Otis, Christian, & Turmel, Monique. 2007. A clade uniting the green algae *Mesostigma viride* and *Chlorokybus atmophyticus* represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. *BMC biology*, **5**(1), 2.

Letsch, M.R., & Lewis, L.A. 2012. Chloroplast gene arrangement variation within a closely related group of green algae (Trebouxiophyceae, Chlorophyta). *Molecular phylogenetics and evolution*, **64**(3), 524–532.

Li, B., Lopes, J.S., F., P.G., Embley, T.M., & Cox, C.J. 2014. Compositional biases among synonymous substitutions cause conflict between gene and protein trees for plastid origins. *Molecular biology and evolution*, **31**(7), 1697–1709.

Liu, Y., Medina, R., & Goffinet, B. 2014. 350 My of mitochondrial genome stasis in mosses, an early land plant lineage. *Molecular biology and evolution*, **31**(10), 2586–2591.

Lombard, V., Bernard, T., Rancurel, C., Brumer, H., Coutinho, P., & Henrissat, B. 2010. A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochemical journal*, **432**, 437–444.

Mach, J. 2011. Cool as the cucumber mitochondrial genome: complete sequencing reveals dynamics of recombination, sequence transfer, and multichromosomal structure. *The Plant Cell Online*, **23**(7), 2472–2472.

Mardis, E.R. 2008. Next-generation DNA sequencing methods. *Annual review of genomics and human genetetics*, **9**, 387–402.

Mardis, E.R. 2011. A decade's perspective on DNA sequencing technology. *Nature*, **470**(7333), 198–203.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y., Chen, Z., & otros. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**(7057), 376–380.

Margulis, L., & Barreno, E. 2003. Looking at lichens. *BioScience*, **53**(8), 776–778.

Marron, A.O., Akam, M., & Walker, G. 2012. Nitrile hydratase genes are present in multiple eukaryotic supergroups. *PLoS One*, **7**(4), e32867.

Matsuzaki, M., Misumi, O., Shin-i, T., Maruyama, S., Takahara, M., Miyagishima, S., Mori, T., & otros. 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature*, **428**(6983), 653–657.

McGinn, N., Yin, Y., Ekstrom, A., & Taujale, R. 2014. PlantCAZyme: a database for plant carbohydrate-active enzymes. *Database*.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Maréchal-Drouard, L., & otros. 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science*, **318**(5848), 245.

Metzker, M.L. 2005. Emerging technologies in DNA sequencing. *Genome research*, **15**(12), 1767–1776.

Michel, F., Costa, M., & Westhof, E. 2009. The ribozyme core of group II introns: a structure in want of partners. *Trends in biochemical sciences*, **34**(4), 189–199.

Miller, J.R., Koren, S., & Sutton, G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics*, **95**(6), 315–327.

Moya, A., Gil, R., Latorre, A., Peretó, J., Garcillán-Barcia, M.P., & De La Cruz, F. 2009. Toward minimal bacterial cells: evolution vs. design. *FEMS Microbiology Reviews*, **33**(1), 225–235.

Muggia, L., Vancurova, L., Škaloud, P.l, Peksa, O., Wedin, M., & Grube, M. 2013. The symbiotic playground of lichen thalli–a highly flexible photobiont association in rock-inhabiting lichens. *FEMS Msicrobiology Ecology*, **85**(2), 313–323.

Nedelcu, A.M., Lee, R.W., Lemieux, C., Gray, M.W., & Burger, G. 2000. The complete mitochondrial DNA sequence of *Scenedesmus obliquus* reflects an intermediate stage in the evolution of the green algal mitochondrial genome. *Genome research*, **10**(6), 819.

Niyogi, K.K., Li, X., Rosenberg, V., & Jung, H. 2005. Is PsbS the site of non-photochemical quenching in photosynthesis? *Journal of Experimental Botany*, **56**(411), 375–382.

Noctor, G., & Foyer, C.H. 1998. Ascorbate and glutathione: keeping active oxygen under control. *Annual review of plant biology*, **49**(1), 249–279.

Odintsova, M.S., & Yurina, N.P. 2005. Genomics and evolution of cellular organelles. *Russian Journal of Genetics*, **41**(9), 957–967.

Orsini, M., Costelli, C., Malavasi, V., Cusano, R., Concas, A., Angius, A., & Cao, G. 2014. Complete genome sequence of mitochondrial DNA (mtDNA) of *Chlorella sorokiniana*. *Mitochondrial DNA*, 1–3.

Ouyang, L., Chen, S., Li, Y., & Zhou, Z. 2013. Transcriptome analysis reveals unique C4-like photosynthesis and oil body formation in an arachidonic acid-rich microalga *Myrmecia incisa* Reisigl H4301. *BMC genomics*, **14**(1), 396.

Palenik, B., Grimwood, J., Aerts, A., Rouzé, P., Salamov, A., Putnam, N., Dupont, C., Jorgensen, R., Derelle, E., Rombauts, S., & otros. 2007. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences*, **104**(18), 7705.

Palmer, Jeffrey D, & Herbon, Laura A. 1988. Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. *Journal of Molecular Evolution*, **28**(1-2), 87–97.

Palmqvist, K. 1993. Photosynthetic CO2-use efficiency in lichens and their isolated photobionts: the possible role of a CO2-concentrating mechanism. *Planta*, **191**(1), 48–56.

Palmqvist, K. 2000. Tansley Review No. 117. Carbon Economy in Lichens. *New Phytologist*, **148**(1), 11–36.

Palmqvist, K., Dahlman, L., Valladares, F., Tehler, A., Sancho, L.G., & Mattsson, J. 2002. CO2 exchange and thallus nitrogen across 75 contrasting lichen associations from different climate zones. *Oecologia*, **133**(3), 295–306.

Palmqvist, K., Dahlman, L., Jonsson, A., & Nash, TH. 2008. The carbon economy of lichens. *Chap. 10, pages 182–215 of:* Nash III, T.H. (ed), *Lichen Biology*, 2 edn. Cambridge UK: Cambridge University Press.

Park, B.H, Karpinets, T.V., Syed, M.H., Leuze, M.R., & Uberbacher, E.C. 2010. CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. *Glycobiology*, **20**(12), 1574–1584.

Parra, G., Bradnam, K., Ning, Z., Keane, T., & Korf, I. 2009. Assessing the gene space in draft genomes. *Nucleic acids research*, **37**(1), 289–297.

Peksa, O., & Skaloud, P. 2008. Changes in chloroplast structure in lichenized algae. *Symbiosis*, **46**(3), 153–160.

Peksa, O., & Škaloud, P. 2011. Do photobionts influence the ecology of lichens? A case study of environmental preferences in symbiotic green alga *Asterochloris* (Trebouxiophyceae). *Molecular ecology*, **20**(18), 3936–3948.

Petrzik, K., Vondrák, J., Kvíderová, J., & Lukavskỳ, J. 2015. Platinum Anniversary: Virus and Lichen Alga Together More than 70 Years. *PLoS One*, **10**(3), e0120768.

Piercey-Normore, M.D. 2006. The lichen-forming ascomycete *Evernia mesomorpha* associates with multiple genotypes of *Trebouxia jamesii*. *New Phytologist*, **169**(2), 331–344.

Pintado, A., Sancho, L.G., Green, T.G., Blanquer, J.M., & Lázaro, R. 2005. Functional ecology of the biological soil crust in semiarid SE Spain: sun and shade populations of *Diploschistes diacapsis* (Ach.) Lumbsch. *The Lichenologist*, **37**(05), 425–432.

Pombert, Jean-François, & Keeling, Patrick J. 2010. The mitochondrial genome of the entomoparasitic green alga Helicosporidium. *PLoS One*, **5**(1), e8954.

Pop, M. 2009. Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics*, **10**(4), 354–366.

Pop, M., Salzberg, S.L., & Shumway, M. 2002. Genome sequence assembly: Algorithms and issues. *Computer*, **35**(7), 47–54.

Prochnik, S.E., Umen, J., Nedelcu, A.M., Hallmann, A., Miller, S.M., Nishii, I., Ferris, P., Kuo, A., Mitros, T., Fritz-Laylin, L.K., & otros. 2010. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science*, **329**(5988), 223.

Ronaghi, M., Uhlén, M., & Nyrén, P. 1998. A sequencing method based on real-time pyrophosphate. *Science*, **281**(5375), 363–365.

Ruhfel, B.R., Gitzendanner, M.A., Soltis, P.S., Soltis, D.E., & Burleigh, J.G. 2014. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC evolutionary biology*, **14**(1), 23.

Sancho, L.G., De la Torre, R., Horneck, G., Ascaso, C., de los Rios, A., Pintado, A., Wierzchos, J., & Schuster, M. 2007. Lichens survive in space: results from the 2005 LICHENS experiment. *Astrobiology*, **7**(3), 443–454.

Sanger, F., Nicklen, S., & Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, **74**(12), 5463–5467.

Schwendener, S. 1869. Die Algentypen der Flechtengonidien. *In: Programm für die Rectorsfeier der Universität Basel*.

Servín-Garcidueñas, L.E., & Martínez-Romero, E. 2012. Complete mitochondrial and plastid genomes of the green microalga Trebouxiophyceae sp. strain MX-AZ01 isolated from a highly acidic geothermal lake. *Eukaryotic cell*, **11**(11), 1417–1418.

Škaloud, P., & Peksa, O. 2010. Evolutionary inferences based on ITS rDNA and actin sequences reveal extensive diversity of the common lichen alga *Asterochloris* (Trebouxiophyceae, Chlorophyta). *Molecular phylogenetics and evolution*, **54**(1), 36–46.

Sloan, D.B., Alverson, A.J., Štorchová, H., Palmer, J.D., & Taylor, D.R. 2010. Extensive loss of translational genes in the structurally dynamic mitochondrial genome of the angiosperm *Silene latifolia*. *BMC evolutionary biology*, **10**(1), 274.

Slocum, R.D., Ahmadjian, V., & Hildreth, K.C. 1980. Zoosporogenesis in *Trebouxia gelatinosa*: Ultrastructure potential for zoospore release and implications for the lichen association. *The Lichenologist*, **12**(02), 173–187.

Smit, A.F.A., Hubley, R., & Green, P. 1996. *RepeatMasker Open-3.0*.

Smith, D., Lee, R., Cushman, J., Magnuson, J., Tran, D., & Polle, J. 2010. The *Dunaliella salina* organelle genomes: large sequences, inflated with intronic and intergenic DNA. *BMC Plant Biology*, **10**(1), 83.

Smith, D.R., Burki, F., Yamada, T., Grimwood, J., Grigoriev, I.V., Van Etten, J.L., & Keeling, P.J. 2011. The GC-Rich mitochondrial and plastid genomes of the green alga *Coccomyxa* give insight into the evolution of organelle DNA nucleotide landscape. *PLoS One*, **6**(8), e23624.

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B.H., & Hood, L.E. 1986. Fluorescence detection in automated DNA sequence analysis. *Nature*, **321**, 674 – 679.

Staden, R., Beal, K.F., & Bonfield, J.K. 1999. The Staden Package, 1998. *Pages 115–130 of:* Misener, Stephen, & Krawetz, StephenA. (eds), *Bioinformatics Methods and Protocols*. Methods in Molecular Biology, vol. 132. Humana Press.

Stanke, M., & Morgenstern, B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research*, **33**(suppl 2), W465–W467.

Swofford, D.L. 2003. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4.

Takeshita, S. 2001. A taxonomic revision of the genus *Trebouxia* (Trebouxiophyceae, Chlorophyta). *Hikobia*, **13**, 425–455.

Tippery, N.P., Fučíková, K., Lewis, P.O., & Lewis, L.A. 2012. Probing the monophyly of the Sphaeropleales (Chlorophyceae) using data from five genes. *Journal of Phycology*, **48**(6), 1482–1493.

Turmel, M., Otis, C., & Lemieux, C. 1999a. The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes. *Proceedings of the National Academy of Sciences*, **96**(18), 10248–10253.

Turmel, M., Lemieux, C., Burger, G., Lang, B.F., Otis, C., Plante, I., & Gray, M.W. 1999b. The complete mitochondrial DNA sequences of *Nephroselmis olivacea* and *Pedinomonas minor*: two radically different evolutionary patterns within green algae. *The Plant Cell Online*, **11**(9), 1717.

Turmel, M., Otis, C., & Lemieux, C. 2002. The complete mitochondrial DNA sequence of *Mesostigma viride* identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants. *Molecular Biology and Evolution*, **19**(1), 24–38.

Turmel, M., Otis, C., & Lemieux, C. 2006. The chloroplast genome sequence of *Chara vulgaris* sheds new light into the closest green algal relatives of land plants. *Molecular Biology and Evolution*, **23**(6), 1324–1338.

Turmel, M., Otis, C., & Lemieux, C. 2007. An unexpectedly large and loosely packed mitochondrial genome in the charophycean green alga *Chlorokybus atmophyticus*. *BMC Genomics*, **8**(1), 137.

Turmel, M., Otis, C., & Lemieux, C. 2009. The chloroplast genomes of the green algae *Pedinomonas minor*, *Parachlorella kessleri*, and *Oocystis solitaria* reveal a shared ancestry between the Pedinomonadales and Chlorellales. *Molecular biology and evolution*, **26**(10), 2317.

Turmel, M., Otis, C., & Lemieux, C. 2013. Tracing the evolution of streptophyte algae and their mitochondrial genome. *Genome biology and evolution*, **5**(10), 1817–1835.

Unseld, M, Marienfeld, Joachim R, Brandt, Petra, & Brennicke, Axel. 1997. The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. *Nature genetics*, **15**, 57–61.

Vahrenholz, C., Riemen, G.and Pratje, E., Dujon, B., & Michaelis, G. 1993. Mitochondrial DNA of *Chlamydomonas reinhardtii*: the structure of the ends of the linear 15.8-kb genome suggests mechanisms for DNA replication. *Current genetics*, **24**(3), 241–247.

Wakasugi, T., Nagai, T., Kapoor, M., Sugita, M., Ito, M., Ito, S., Tsudzuki, J., Nakashima, K., Tsudzuki, T., Suzuki, Y., & otros. 1997. Complete nucleotide sequence of the chloroplast genome from the green alga *Chlorella vulgaris*: the existence of genes possibly involved in chloroplast division. *Proceedings of the National Academy of Sciences*, **94**(11), 5967.

Wang, Y., Zhang, T., Zhou, Q., & Wei, J. 2011. Construction and characterization of a full-length cDNA library from mycobiont of *Endocarpon pusillum* (lichen-forming Ascomycota). *World Journal of Microbiology and Biotechnology*, **27**(12), 2873–2884.

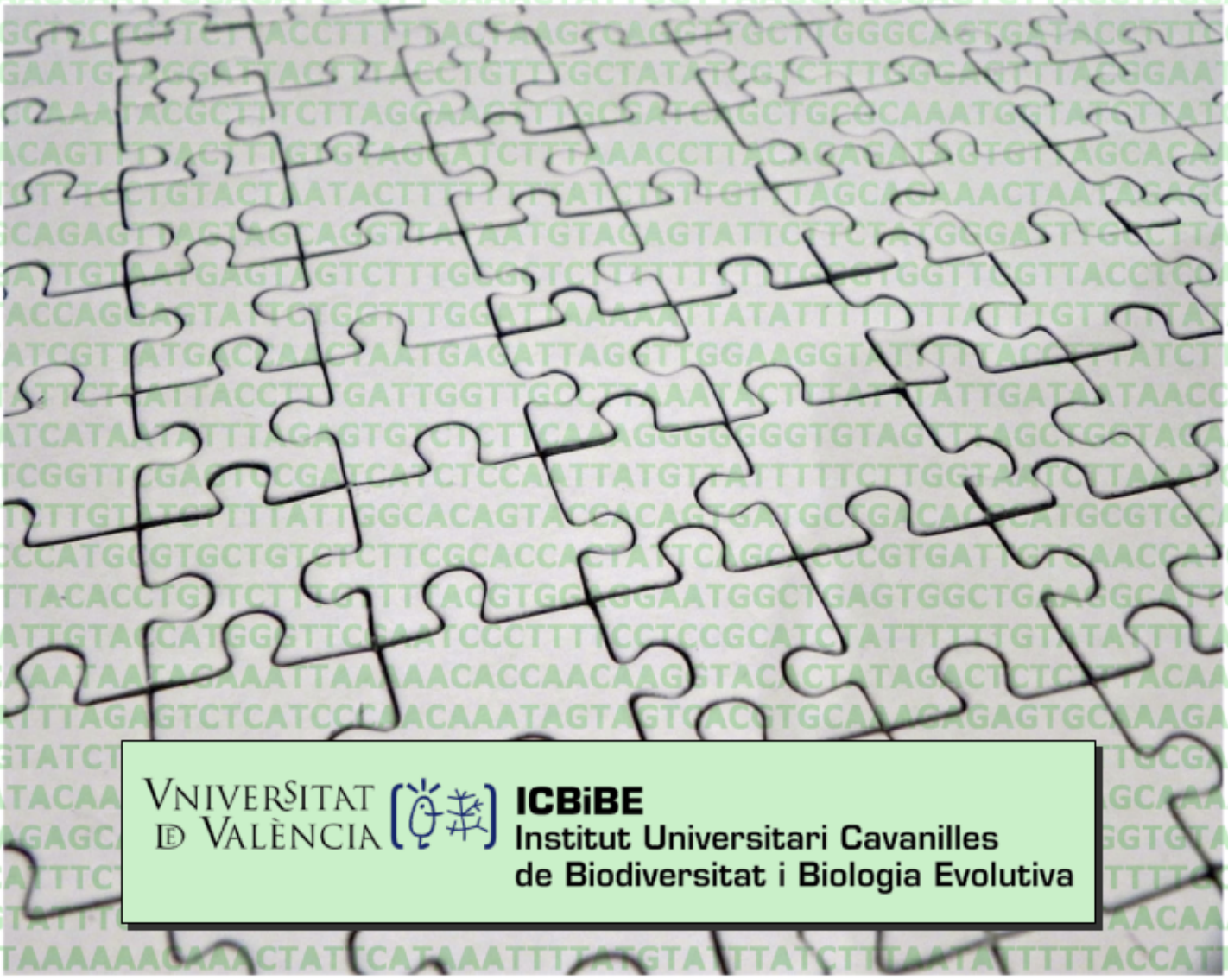
Wolff, G., Plante, I., Lang, B.F., KÃŒck, U., & Burger, G. 1994. Complete sequence of the mitochondrial DNA of the chlorophyte alga *Prototheca wickerhamii*: Gene content and genome organization. *Journal of molecular biology*, **237**(1), 75–86.

Worden, A.Z., Lee, J.H., Mock, T., Rouzé, P., Simmons, M.P., Aerts, A.L., Allen, A.E., Cuvelier, M.L., Derelle, E., Everett, M.V., & otros. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science*, **324**(5924), 268.

Yandell, M., & Ence, D. 2012. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, **13**(5), 329–342.

Zdobnov, E.M., & Apweiler, R. 2001. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**(9), 847–848.

Zerbino, D.R., & Birney, E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, **18**(5), 821–829.



VNIVERSITAT DE VALÈNCIA

**ICBiBE**
**Institut Universitari Cavanilles de Biodiversitat i Biologia Evolutiva**